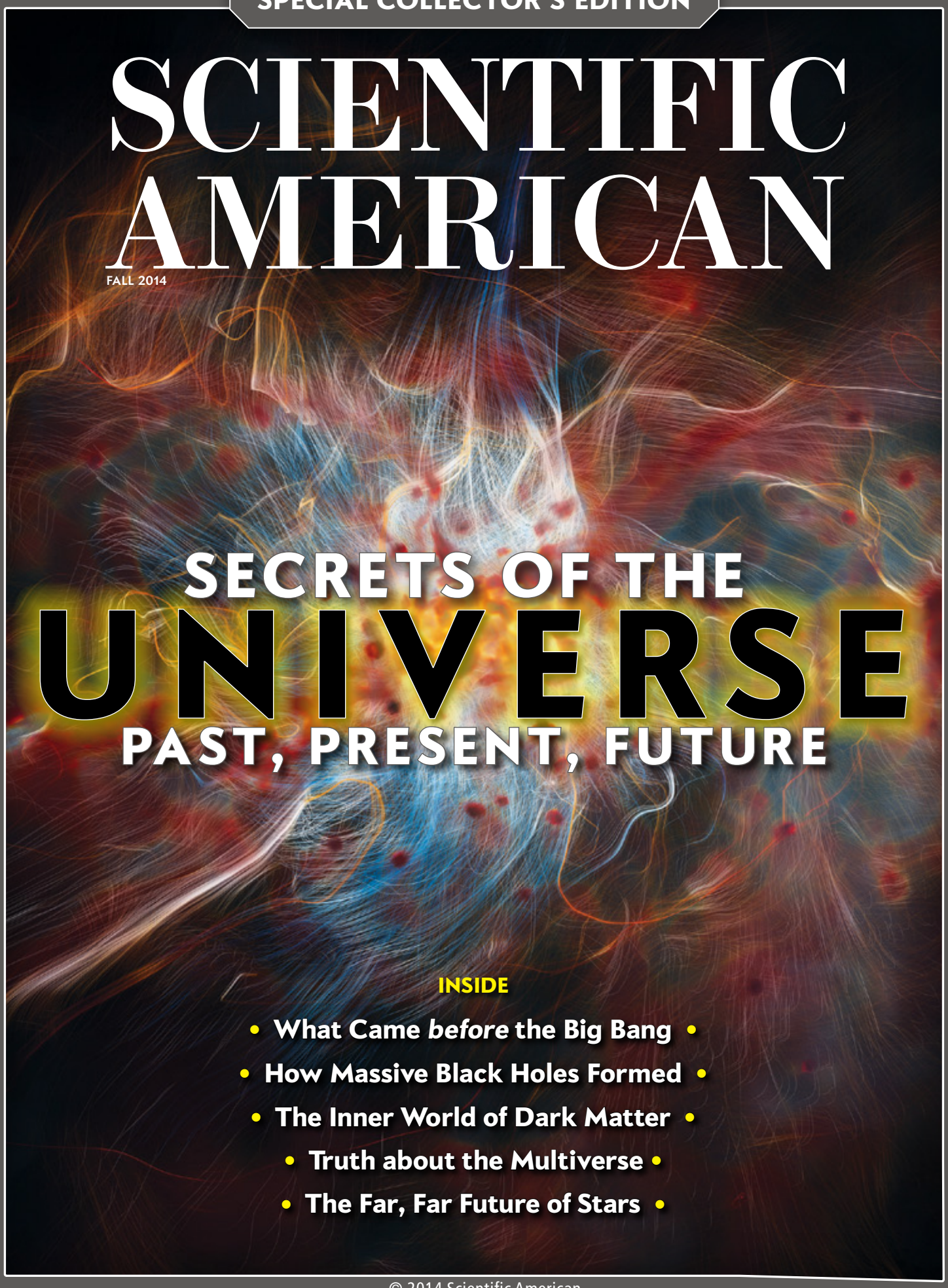
SPECIAL COLLECTOR'S EDITION

SCIENTIFIC AMERICAN

FALL 2014



SECRETS OF THE UNIVERSE

PAST, PRESENT, FUTURE

INSIDE

- What Came *before* the Big Bang
- How Massive Black Holes Formed
- The Inner World of Dark Matter
- Truth about the Multiverse
- The Far, Far Future of Stars

© 2014 Scientific American

Get All-Access to THE ARCHIVES

EVERY ISSUE. EVERY YEAR. 1845–PRESENT.

All-Access Subscriptions Now Available!

PRINT
12 issues featuring the latest developments in science, medicine and technology.

+

ARCHIVES
Search over 4,960 issues and 133,700 articles to find the information you need. 1845–present.

+

DIGITAL
Instant access to current issues and the archive using any computer, tablet or smartphone.*

www.scientificamerican.com/all-access

SCIENTIFIC AMERICAN™

*Internet connection required. Copyright © 2014 by Scientific American, a division of Nature America, Inc. All right reserved.

SCIENTIFIC AMERICAN™
ESTABLISHED 1845

Secrets of the Universe is published by the staff of Scientific American, with project management by:

EDITOR IN CHIEF AND SENIOR VICE PRESIDENT
Mariette DiChristina
EXECUTIVE EDITOR **Fred Guterl**
MANAGING EDITOR **Ricki L. Rusting**
ISSUE EDITOR **Lee Billings**

DESIGN DIRECTOR **Michael Mrak**
ISSUE DESIGNER **Lawrence R. Gendron**
PHOTOGRAPHY EDITOR **Monica Bradley**
PHOTO RESEARCHER **Liz Tormes**
MANAGING PRODUCTION EDITOR **Richard Hunt**
SENIOR PRODUCTION EDITOR **Michelle Wright**

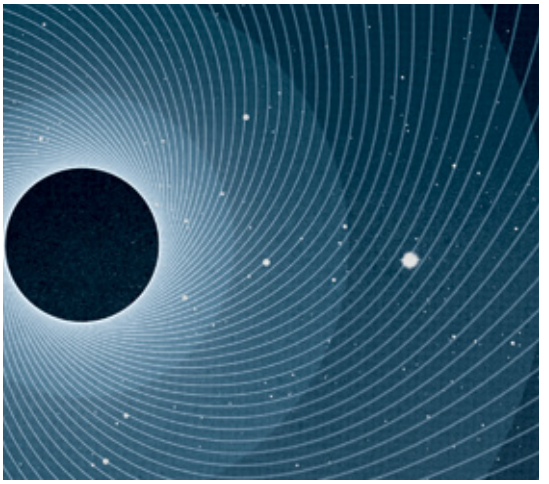
COPY DIRECTOR **Maria-Christina Keller**
SENIOR COPY EDITOR **Daniel C. Schlenoff**
COPY EDITOR **Aaron Shattuck**

EDITORIAL ADMINISTRATOR **Avonelle Wing**
SENIOR SECRETARY **Maya Harty**

SENIOR PRODUCTION MANAGER
Christina Hippeli
ADVERTISING PRODUCTION MANAGER
Carl Cherebin
PREPRESS AND QUALITY MANAGER
Silvia De Santis
CUSTOM PUBLISHING MANAGER
Madelyn Keyes-Milch
PRODUCTION COORDINATOR **Lisa Headley**

PRESIDENT
Steven Inchcoombe
EXECUTIVE VICE PRESIDENT
Michael Florek
VICE PRESIDENT AND ASSOCIATE PUBLISHER, MARKETING AND BUSINESS DEVELOPMENT
Michael Voss
DIRECTOR, INTEGRATED MEDIA SALES
Stan Schmidt
ASSOCIATE VICE PRESIDENT, BUSINESS DEVELOPMENT
Diane McGarvey
DIRECTOR, GLOBAL MEDIA SOLUTIONS
Jeremy A. Abbate
VICE PRESIDENT, CONSUMER MARKETING
Christian Dorbandt
DIRECTOR, E-COMMERCE **Scott Rademaker**
ASSOCIATE CONSUMER MARKETING DIRECTOR
Catherine Bussey
E-COMMERCE MARKETING MANAGER
Evelyn Veras
SENIOR MARKETING MANAGER / ACQUISITION
Patricia Elliott
ONLINE MARKETING ASSOCIATE
Alexandra Egan
MARKETING AND CUSTOMER SERVICE COORDINATOR
Christine Kaelin
SALES DEVELOPMENT MANAGER **David Tirpack**
PROMOTION MANAGER **Diane Schube**
PROMOTION ART DIRECTOR **Maria Cruz-Lord**
MARKETING RESEARCH DIRECTOR **Rick Simone**
ONLINE MARKETING PRODUCT MANAGER
Zoya Lysak
CORPORATE PR MANAGER **Rachel Scheer**
SENIOR INTEGRATED SALES MANAGER
Jay Berfas
SALES REPRESENTATIVE **Chantel Arroyo**
INTEGRATED MEDIA SALES **Nate Murray**
SENIOR ADMINISTRATOR, EXECUTIVE SERVICES
May Jung
CUSTOM PUBLISHING EDITOR **Lisa Pallatroni**
RIGHTS AND PERMISSIONS MANAGER
Karin M. Tucker

GAVIN POTENZA; SCIENTIFIC AMERICAN IS A TRADEMARK OF SCIENTIFIC AMERICAN, INC., USED WITH PERMISSION



Answers in the Sky

When we gaze into the peaceful, twinkling night sky, we see much the same vistas witnessed by our ancestors thousands of years ago. Yet centuries of scientific progress have so transformed our understanding that it cannot be said we share their same simplistic view. We now know that the world is not flat, that the sky is not a dome, that the stars are not pinholes through which flickering fires shine.

Many mysteries remain, of course. But with every year, new discoveries bring a better understanding of how our cosmos emerged, how it evolved and where it is going. Today we know that the stars are suns and that Earth is just one of billions of planets in our Milky Way galaxy. Scientists are racing to learn whether or not we are cosmically alone, looking for Earth-like planets and studying them for signs of life.

Life here on Earth can be traced back to events more than 4.6 billion years ago, when our sun was born as part of a larger brood in a cluster of stars. Some astronomers are on the verge of tracking down our long-lost starry siblings and reconstructing the solar system's deepest history. Meanwhile others are exploring the possibilities for life in the distant future, attempting to glimpse the universe hundreds of billions of years from now.

On larger scales, the universe offers other surprises of great importance for us. Galaxies are held together by the invisible, gravitational glue of dark matter and are moving away from one another at an accelerating pace that is the result of a force known only as dark energy. A similar force may have been active a trillionth of a second after the big bang, when the universe's expansion briefly accelerated in a process called primordial inflation. If the properties of dark matter, dark energy or inflation had been slightly different, galaxies and stars could not have formed, precluding the possibility of life as we know it.

These discoveries are pushing science into unexplored frontiers. Many cosmologists believe that inflation led not only to our entire observable universe but also to an infinite "multiverse" of parallel universes. And a few maverick researchers are investigating the idea of a cyclic universe, a cosmos eternally oscillating between death and rebirth. Physicists are also studying whether information, rather than matter and energy, forms the fabric of spacetime.

The articles in this special edition of *Scientific American* delve into all of these mind-bending ideas and more. The night sky will never seem simple again.

Lee Billings
Issue Editor
editors@sciam.com

© 2014 Scientific American

SPECIAL EDITION

Secrets of the Universe

28

1 **Introduction**

PLANETS AND STARS

4 **The Dawn of Distant Skies**
The galaxy is teeming with planets. Scientists are straining to peer into their atmospheres for signs of extraterrestrial life.
By Michael D. Lemonick

12 **The Long-Lost Siblings of the Sun**
The sun was born in a family of stars. What became of them?
By Simon F. Portegies Zwart

20 **The Inner Life of Star Clusters**
All stars arise in groups but then slowly disperse into space. A new theory seeks to explain how these groups form and fall apart or, in rare cases, persist for hundreds of millions of years.
By Steven W. Stahler

28 **The Far, Far Future of Stars**
Some say its glory days are long gone, but the universe has life in it yet. Brand-new types of celestial phenomena will unfold over the coming billions and trillions of years.
By Donald Goldsmith

THE DARK COSMOS

36 **Goldilocks Black Holes**
Tipping the scales at less than about a million suns in mass, middleweight black holes hold clues to how their much larger siblings, and galaxies, first formed.
By Jenny E. Greene

44 **Dark Worlds**
A shadow cosmos, woven silently into ours, may have its own rich inner life.
By Jonathan Feng and Mark Trodden

52 **The Lost Galaxies**
By the latest estimate, the observable universe contains 200 billion galaxies. Astronomers wonder: Why so few?
By James E. Geach

60 **The Universe's Invisible Hand**
Dark energy does more than hurry along the expansion of the universe. It also has a stranglehold on the shape and spacing of galaxies.
By Christopher J. Conselice

ON THE COVER

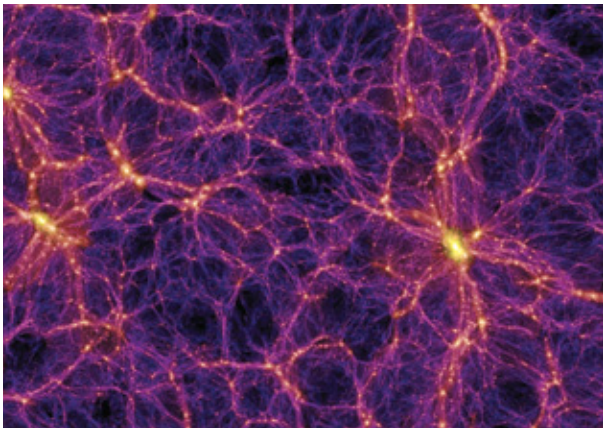
In this state-of-the-art simulation of galaxy formation, a galaxy like our Milky Way pulls in red streams of cold, dense gas and ejects hot, blue gas into intergalactic space. Only a small fraction of material remains within the galaxy. Image by James E. Geach and Rob Crain.

© 2014 Scientific American

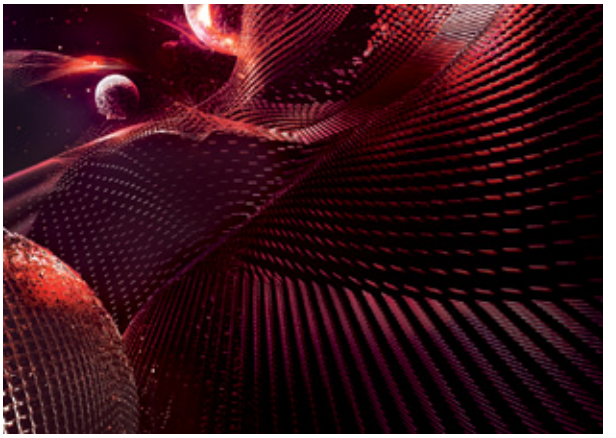
SCIENTIFIC AMERICAN

Volume 23, No. 3, Fall 2014

20 Most stars, including our sun, formed in clusters.



44 Dark matter and dark energy sculpt cosmic structure.



104 Our universe may be digital, made up of information.

THE UNIVERSE

THE EDGE OF EXISTENCE

68 The Inflation Debate
Is the theory at the heart of modern cosmology deeply flawed?
By Paul J. Steinhardt

76 Does the Multiverse Really Exist?
Proof of parallel universes radically different from our own may lie beyond the domain of science.
By George F. R. Ellis

82 Follow the Bouncing Universe
Our universe could have started not with a big bang but with a big bounce—an implosion that triggered an explosion, all driven by exotic quantum-gravitational effects.
By Martin Bojowald

THE NATURE OF SPACE AND TIME

88 Is the Universe Leaking Energy?
Total energy must be conserved. Every student of physics learns this fundamental law. The trouble is, it does not apply to the universe as a whole.
By Tamara M. Davis

96 A Geometric Theory of Everything
Deep down, the particles and forces of the universe are a manifestation of exquisite geometry.
By A. Garrett Lisi and James Owen Weatherall

104 Is Space Digital?
An experiment going up outside of Chicago will attempt to measure the intimate connections among information, matter and spacetime. If it works, it could rewrite the rules for 21st-century physics.
By Michael Moyer

112 The Longest View
To learn how the universe evolved over time, a space telescope gazes back to the earliest galaxies ever observed.
By Lee Billings

Articles in this special issue are updated from previous issues of *Scientific American*.

Scientific American Special (ISSN 1936-1513), Volume 23, Number 3, Fall 2014, published by Scientific American, a division of Nature America, Inc., 75 Varick Street, 9th Floor, New York, NY 10013-1917. Canadian BN No. 127387652RT; TVQ1218059275 TQ0001. To purchase additional quantities: U.S., $13.95 each; elsewhere, $17.95 each. Send payment to Scientific American Back Issues, 75 Varick Street, 9th Floor, New York, NY 10013-1917. **Inquiries: fax 212-355-0408 or telephone 212-451-8415. Printed in U.S.A.**

Copyright © 2014 Scientific American, a division of Nature America, Inc. All rights reserved.

© 2014 Scientific American

ASTRONOMY

THE DAWN OF DISTANT SKIES

The galaxy is teeming with planets. Scientists are straining to peer into

© 2014 Scientific American

their atmospheres to seek signs of extraterrestrial life

By Michael D. Lemonick

© 2014 Scientific American

NOBODY WHO WAS THERE AT THE TIME, FROM THE MOST SEASONED ASTROPHYSICIST to the most inexperienced science reporter, is likely to forget a press conference at the American Astronomical Society's winter meeting in San Antonio, Tex., in January 1996. It was there that Geoffrey W. Marcy, an observer then at San Francisco State University, announced that he and his observing partner, R. Paul Butler, then at the University of California, Berkeley, had discovered the second and third planets ever found orbiting a sunlike star. The first such planet, 51 Pegasi b, had been announced by Michel Mayor and Didier Queloz of the University of Geneva a few months earlier—but a single detection could have been a fluke or even a mistake. Now Marcy was able to say confidently that it had been neither. "Planets," he told the crowd, "aren't rare after all."

The announcement shook the world of astronomy. Almost nobody had been looking for planets because scientists were convinced they would be too hard to find. Now, after searching a mere handful of stars, astronomers had discovered three worlds, suggesting billions more waiting to be found.

If Butler and Marcy had merely settled a question on planetary formation theory, their discovery would not have been such a big deal. But it showed unequivocally that so-called extrasolar planets did exist and, with them, the possibility of answering a question that had vexed philosophers, scientists and theologians since the time of the ancient Greeks: Are we alone in the universe?

After the initial celebration, scientists settled down to figuring out exactly how they were going to investigate the prospect of even a rudimentary form of life on a planet orbiting an alien sun. Short of picking up an extraterrestrial broadcast, à la Jodie Foster in the movie *Contact,* the only way to find out would be to search extrasolar planets for atmospheric biosignatures—evidence of highly reactive molecules such as oxygen that would quickly disappear unless some kind of metabolizing organisms were replenishing the supply.

Marcy, Mayor and their colleagues had seen only the gravitational effect the planets had on their parent star; to detect a biosignature, you would need to image an exoatmosphere directly.

IN BRIEF

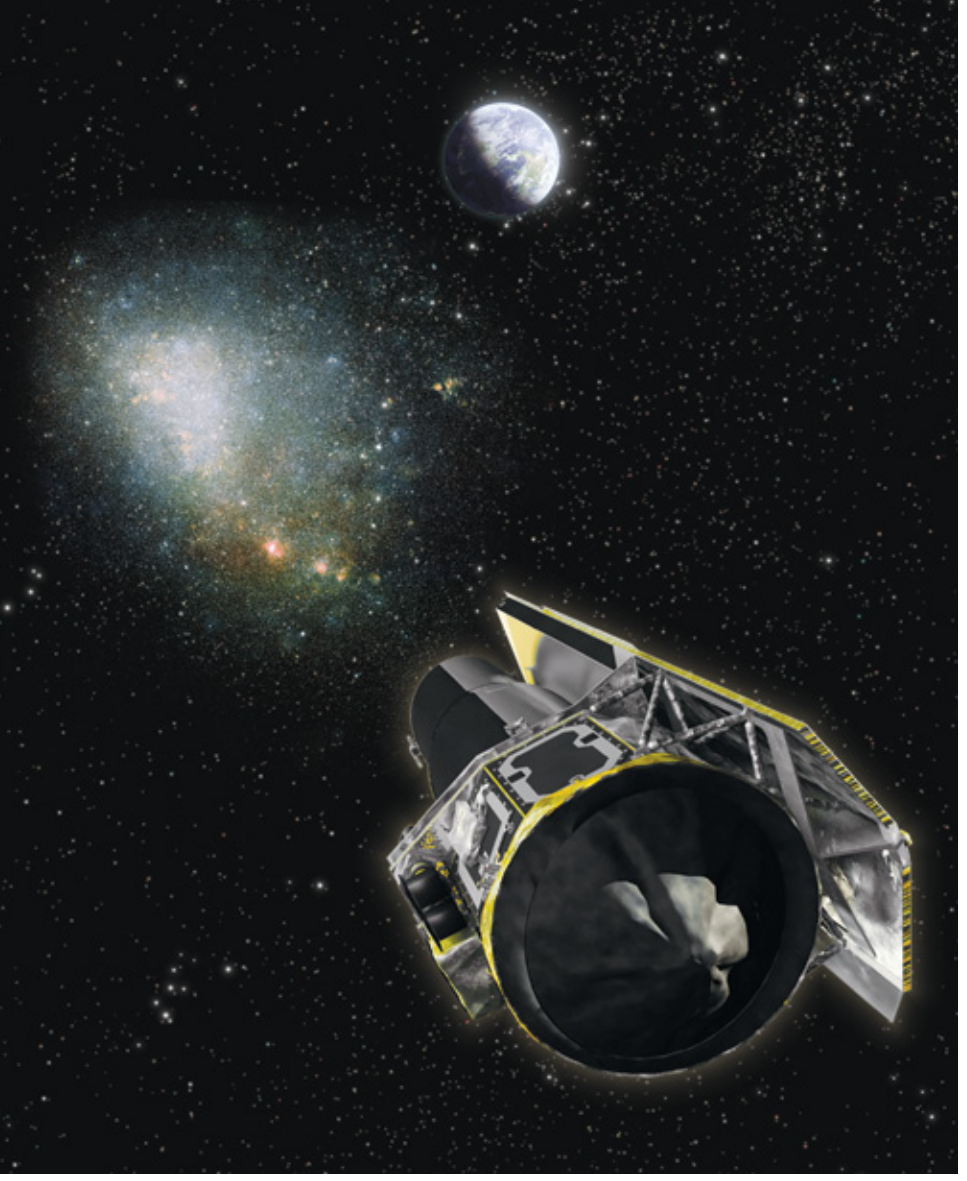
Common wisdom once held that it would be nearly impossible to investigate the atmospheres of distant exoplanets—the glare from their parent star would be too bright.

Yet once scientists began to study exoplanets as they passed behind their star, they realized that the resulting change in stellar brightness could provide clues to what the atmospheres are made of.

Astronomers are now using these advanced techniques to detect atoms and molecules of exoplanetary atmospheres. They hope to soon extend their search to molecules that will provide evidence of distant life.

PRECEDING PAGES: COURTESY OF EUROPEAN SOUTHERN OBSERVATORY/L. CALÇADA

© 2014 Scientific American



SOLAR ECLIPSE: The Spitzer Space Telescope can detect the minute change in brightness that happens when a planet passes behind its host star.

COURTESY OF NASA/CALTECH

To do this, NASA planned to launch an increasingly powerful series of space telescopes, a program that would culminate in an orbiting telescope called the Terrestrial Planet Finder Interferometer that would cost billions of dollars and fly sometime in the 2020s. In short, astronomers knew that they wouldn't be learning anything about exoplanet atmospheres anytime soon.

They were wrong. The discovery of those first few exoplanets inspired an entire generation of young scientists to get into what was suddenly the hottest specialty in astrophysics. It convinced many of their older colleagues to switch into exoplanetology as well. This sudden influx of brainpower led to fresh ideas for investigating exoplanet atmospheres and sped things up dramatically. By 2001 observers had identified sodium in the atmosphere of one exoplanet. Since then, they have identified methane, carbon dioxide, carbon monoxide and water as well. They have even found indirect hints, by examining exoplanet atmospheres, that some planets may be partly made of pure diamond. "At this point," says Heather Knutson, a California Institute of Technology astrophysicist who was involved in many of these pioneering observations, "we've learned something about the atmospheres on the order of 30 to 50 planets—if you count stuff that's not yet published."

These discoveries are still a long way from providing evidence of life—no surprise, since most of the worlds Knutson is talking about are hot, Jupiter-like planets that hug their star more tightly than fiery Mercury orbits the sun. Increasingly, however, Knutson and other observers have begun to probe the atmospheres of smaller planets, so-called super-Earths, which are between two and 10 times as massive as our home planet—something that nobody could have imagined just a decade ago. The announcement in April 2013 that the Kepler space telescope had found two planets less than twice Earth's size, both in orbits where temperatures might permit life to survive, was one more hint that life-friendly worlds are almost certainly plentiful. So while these planets, named Kepler 62e and 62f, are too distant to study in detail, astronomers are convinced it will not be many more years before observers can look for biosignatures in the atmospheres of planets that are essentially twins of Earth.

THE PARKING-LOT PLANET

ASTRONOMERS assumed it would take decades to start looking at planetary atmospheres because the first handful of exoplanets were discovered indirectly, through the influence each had on its parent star. The planets themselves were invisible, but because each star and planet orbit a mutual center of gravity, the gravitational tug of the planet makes the star appear to wobble in place. When a star moves toward us, its light subtly shifts toward the blue end of the visible-light spectrum; when it moves away, the light shifts to the red. The degree of shifting tells observers the star's radial velocity, or how fast it moves toward and away from Earth, which in turn tells us how massive the exoplanet is.

Another option for finding planets was also available, however. If the invisible planet's orbit were perfectly edge-on as seen from Earth, the planet would pass directly in front of its star, in what is known as a transit. Yet at the time of those first discoveries nearly two decades ago, few astrophysicists were thinking about transits at all, simply because the search for planets was itself so far out on the fringe. (A notable exception was William J. Borucki of the NASA Ames Research Center, whose Kepler spacecraft would eventually find transiting objects by the thousands.)

A few years later, in 1999, Timothy W. Brown, then at the National Center for Atmospheric Research, and David Charbonneau, at the time a graduate student at Harvard University, set up a tiny, amateur-size telescope in a parking lot in Boulder, Colo., and saw an exoplanet transit for the first time. The plan-

© 2014 Scientific American

et was HD 209458b, which had been detected earlier by the radial-velocity technique. Weeks later Gregory W. Henry of Tennessee State University, working with Marcy, watched the same planet transiting its star. Both teams have been given equal credit for the discovery because the two detections were published simultaneously.

The successful detection of transits not only gave astronomers a second way to find exoplanets, it also gave them a way to measure their density. The radial-velocity technique had revealed HD 209458b's mass. Now astronomers knew how physically large it was because the amount of starlight a planet blocks is directly proportional to its size. (Dividing its mass by its size showed HD 209458b to be 38 percent larger than Jupiter even though the planet is only 71 percent as massive, an unexpected result that Princeton University astrophysicist Adam Burrows refers to as "an ongoing problem to explain.")

By this time a number of astrophysicists had realized that transits also made it possible to study an exoplanet's atmosphere, in what Knutson calls a "wonderfully clever shortcut." Even before the first transit was reported, in fact, Sara Seager, an astrophysicist at the Massachusetts Institute of Technology, who at the time was Charbonneau's fellow grad student at Harvard, had co-authored a paper with her adviser, Dimitar D. Sasselov, in which they predicted what an observer should see as light from a star passed through a planet's atmosphere when the planet moved across the star's face [see "Planets We Could Call Home," by Dimitar D. Sasselov and Diana Valencia; Scientific American, August 2010]. Physicists have long known that different atoms and molecules absorb light at different wavelengths. If you look at planets in a wavelength that corresponds to the molecule you are searching for, any atmospheres containing that molecule will absorb the light. The wispy planetary atmosphere will become opaque, making the planet appear larger.

Seager and Sasselov suggested that sodium would be especially easy to detect. "Sodium is like skunk scent," Charbonneau says. "A little bit goes a long way." He knows this better than anyone: in 2001 Charbonneau, Brown and their colleagues went back to HD 209458b, their original transiting planet, not with a puny amateur telescope but with the Hubble Space Telescope. Sure enough, the sodium signal was there, just as predicted.

TOTAL ECLIPSE

Astronomers also realized that there was a second, complementary way to inspect the atmospheres of transiting planets. When a planet passes in front of its star, it presents its nightside to the observer. At other times, it shows at least part of its dayside, and just before the planet goes behind the star, the dayside is facing Earth. Although the star is far, far brighter, the planet itself also glows, mostly in the infrared.

That glow vanishes abruptly, however, when the planet moves behind the star; its contribution to the combined light of planet and star vanishes. If astrophysicists can do a before-and-after comparison, they can deduce what the planet alone would look like [*see box on opposite page*]. "It changes the nature of the problem," Knutson says. "Instead of having to detect a very faint thing close to a very bright thing, all you have to do is measure signals that change with time." As early as 2001, L. Drake Deming, then at the NASA Goddard Space Flight Center, aimed an infrared telescope on Hawaii's Mauna Kea at HD 209458b in an attempt to see this so-called secondary eclipse, but, he says, he couldn't make a detection.

He knew, however, that the Spitzer Space Telescope, scheduled for launch in 2003, would almost certainly be able to make such an observation, as did Charbonneau. Both astrophysicists, unbeknownst to each other, applied for time on Spitzer to make the observations. Both got the time and took their data. Then, one day in early 2005, Deming recalls, he got a voice message: "Drake, this is Dave Charbonneau of Harvard," the voice said. "I hear you made some interesting observations lately. Maybe we should talk."

ASTRONOMERS ARE CONVINCED IT WON'T BE MANY MORE YEARS BEFORE OBSERVERS CAN LOOK FOR BIOSIGNATURES IN THE ATMOSPHERES OF PLANETS THAT ARE ESSENTIALLY TWINS OF EARTH.

It turned out that Deming (working with Seager) and Charbonneau had independently made the first secondary-eclipse detections in history, at virtually the same time, using the same observatory. The two groups announced their results for two different stars—the much worked over HD 209458b in Deming's case and a planet named TrES-1 in Charbonneau's—simultaneously. A year later Deming's team detected the secondary eclipse of a planet called HD 189733b. "This," wrote Seager and Deming in a 2010 review article, "unleashed a flood of secondary eclipse observational detections using Spitzer.... It is accurate to say that no one anticipated the full magnitude and stunning impact of the Spitzer Space Telescope as a tool to develop the field of exoplanet atmospheric studies." In fact, Seager says, "we're using the Hubble and the Spitzer in ways they were never designed to be used, going to decimal places they were never designed to reach."

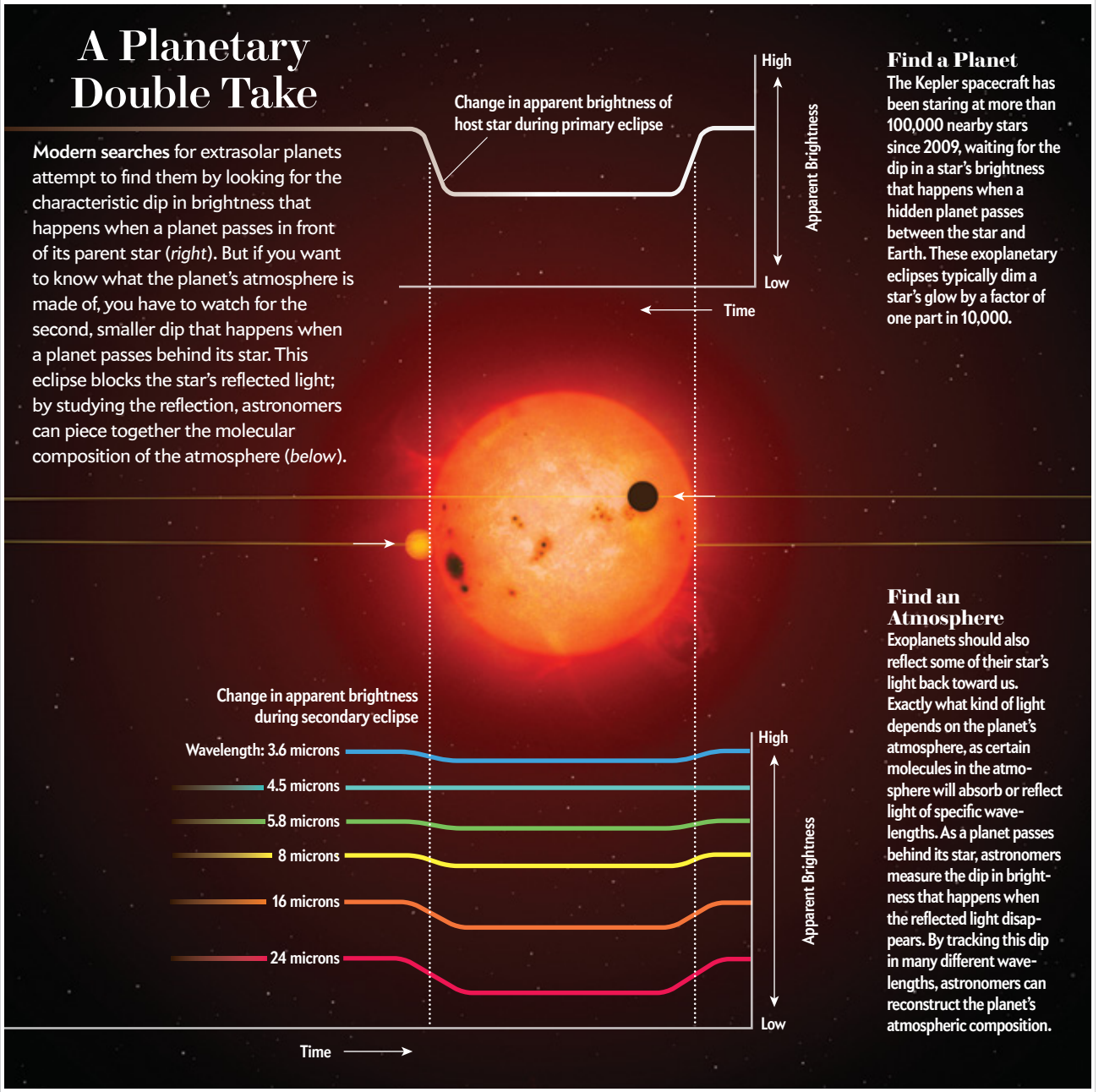
ATMOSPHERIC LAYERS

Those studies have shown a couple of things, Seager says. "This sounds trite in a way, but we've learned that hot Jupiters are hot. We've measured their brightness and temperatures," and what scientists have observed is consistent with how they expect stars to heat their planets. "Number two," she continues, "we've detected molecules. Now, has [what we've found] been very different from what we expected? You know, not really." Seager notes that physicists can straightforwardly model a ball of gas at some temperature made of some combination of elements and ask what kind of molecules form. "The laws of physics and chemistry are universal," she says.

Seager and other astrophysicists have also learned, however, that despite the overall similarity of exoplanet atmospheres, individual planets can differ in several ways. One has to do with how temperature changes with altitude. Some planets, such as Jupiter and Saturn in our own solar system, show temperature inversions, in which temperature rises with altitude rather than falling. Others do not. "The problem," Knutson observes, "is that we don't know what's causing the inversion,

© 2014 Scientific American

HOW IT WORKS


A Planetary Double Take
Modern searches for extrasolar planets attempt to find them by looking for the characteristic dip in brightness that happens when a planet passes in front of its parent star (right). But if you want to know what the planet's atmosphere is made of, you have to watch for the second, smaller dip that happens when a planet passes behind its star. This eclipse blocks the star's reflected light; by studying the reflection, astronomers can piece together the molecular composition of the atmosphere (below).
Change in apparent brightness of host star during primary eclipse
High
Low
Apparent Brightness
Time
Find a Planet
The Kepler spacecraft has been staring at more than 100,000 nearby stars since 2009, waiting for the dip in a star's brightness that happens when a hidden planet passes between the star and Earth. These exoplanetary eclipses typically dim a star's glow by a factor of one part in 10,000.
Change in apparent brightness during secondary eclipse
Wavelength: 3.6 microns
4.5 microns
5.8 microns
8 microns
16 microns
24 microns
High
Low
Apparent Brightness
Time
Find an Atmosphere
Exoplanets should also reflect some of their star's light back toward us. Exactly what kind of light depends on the planet's atmosphere, as certain molecules in the atmosphere will absorb or reflect light of specific wavelengths. As a planet passes behind its star, astronomers measure the dip in brightness that happens when the reflected light disappears. By tracking this dip in many different wavelengths, astronomers can reconstruct the planet's atmospheric composition.

SOURCE: NASA/JPL-CALTECH/K. STEVENSON *University of Central Florida (secondary eclipse brightness data)*

and we can't predict, therefore, which exoplanets will and won't have this feature." Some astrophysicists suggest that exoplanets with inversions might have some kind of heat-absorbing molecule, such as titanium oxide, but so far this is just a hypothesis.

Another question is whether certain planetary atmospheres are made from a different mix of molecules than others. Nikku Madhusudhan, now at the University of Cambridge, analyzed the visible and infrared signature of a planet named WASP-12b and deduced that its atmosphere is unusually rich in carbon, with about as much of that element as oxygen.

Theory suggests that a carbon-to-oxygen ratio of more than 0.8, if mirrored in other, smaller planets in the same system (as it presumably would be, given that planets in a solar system are thought to condense from a single disk of gas and dust), would lead to "rocks" made of carbides—carbon-rich minerals—rather than the silicon-rich silicate rocks found in our solar system. If that were true, an Earth-size planet in the WASP-12 system could have continents made of diamond.

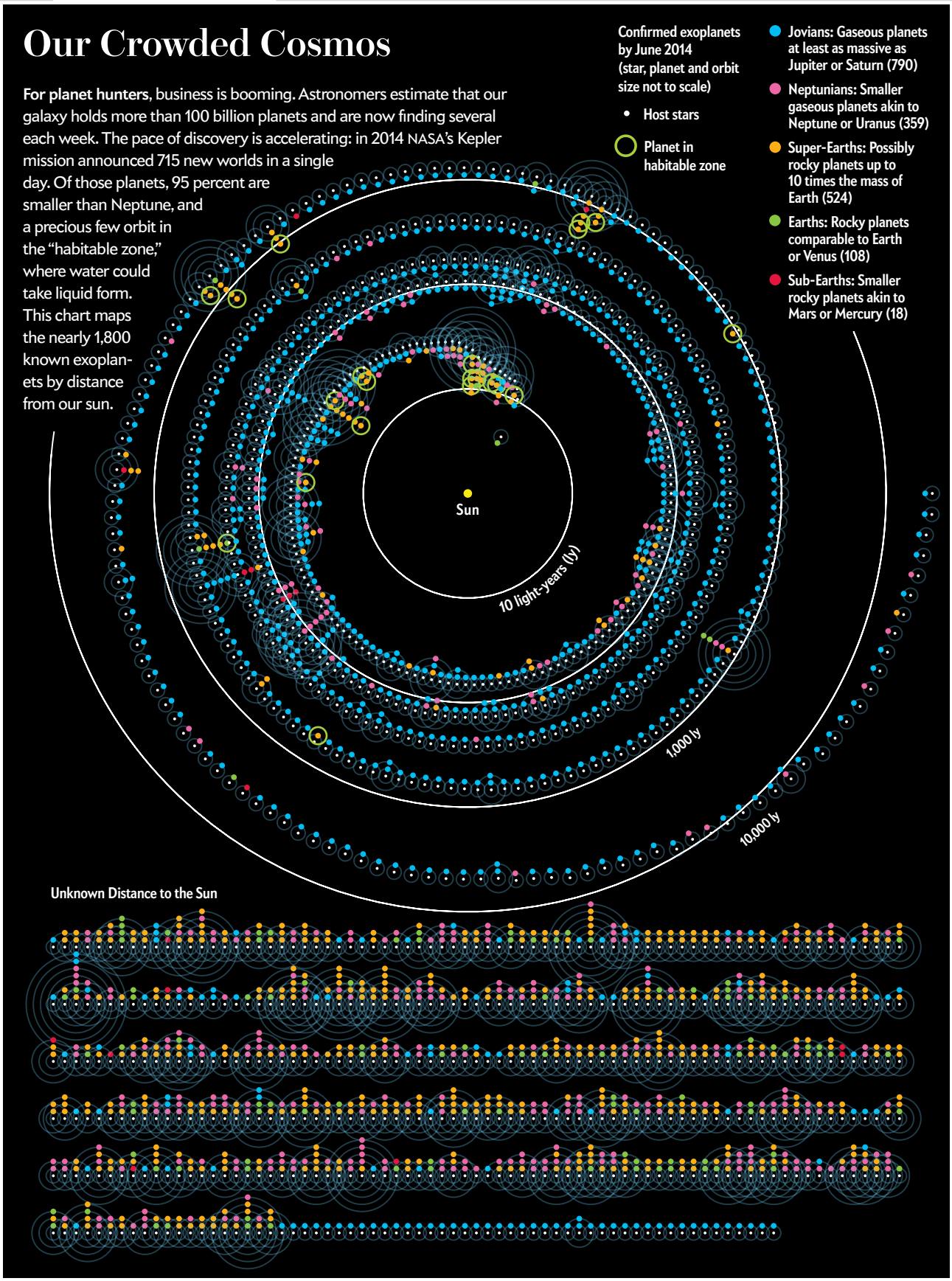
Seager and others have written theoretical papers suggesting that there is nothing to rule out planets made largely of carbon or even of iron. In the case of WASP-12, however, it may not be correct. Knutson says that Ian Crossfield of the Max Planck Institute for Astronomy in Heidelberg, Germany, recently found that the light from WASP-12 is contaminated with light from a fainter double star in the background. "His data perhaps seem

Illustration by Mark A. Garlick

© 2014 Scientific American

Our Crowded Cosmos

For planet hunters, business is booming. Astronomers estimate that our galaxy holds more than 100 billion planets and are now finding several each week. The pace of discovery is accelerating: in 2014 NASA's Kepler mission announced 715 new worlds in a single day. Of those planets, 95 percent are smaller than Neptune, and a precious few orbit in the "habitable zone," where water could take liquid form. This chart maps the nearly 1,800 known exoplanets by distance from our sun.


Confirmed exoplanets by June 2014 (star, planet and orbit size not to scale)
Host stars
Planet in habitable zone
Jovians: Gaseous planets at least as massive as Jupiter or Saturn (790)
Neptunians: Smaller gaseous planets akin to Neptune or Uranus (359)
Super-Earths: Possibly rocky planets up to 10 times the mass of Earth (524)
Earths: Rocky planets comparable to Earth or Venus (108)
Sub-Earths: Smaller rocky planets akin to Mars or Mercury (18)
Sun
10 light-years (ly)
1,000 ly
10,000 ly
Unknown Distance to the Sun

SOURCE: DATA FROM PHL'S EXOPLANET CATALOG OF THE PLANETARY HABITABILITY LABORATORY, UPR ARECIBO

Graphic by Jan Willem Tulp

© 2014 Scientific American

to cast some doubt on the interpretation for this particular planet," Knutson says.

WATER WORLD

BY FAR THE MOST INTENSE FOCUS of observations has been concentrated on a planet known GJ 1214b, which orbits a small, reddish "M-dwarf" star lying about 40 light-years from Earth. Its proximity makes GJ 1214b relatively easy to study, and its size, just 2.7 times the width of Earth, makes it far closer to being Earth-like than the hot Jupiters found in the first years of planet hunting. "It is everybody's favorite super-Earth," says Laura Kreidberg, a grad student at the University of Chicago who is leading the data analysis on one such observing project.

GJ 1214b was found in 2009 during the so-called MEarth Project organized by Charbonneau to look for planets around M dwarfs. The idea was that small transiting planets would be easier to find around these small, dim stars than around bigger ones, for several reasons. First, an Earth-size planet would block a relatively greater percentage of the small star's light. Such a planet would also exert a relatively greater gravitational pull on the star, making it easier to gauge the planet's mass and thus its density. The habitable zone for a small, cool star would also be much closer in than it is for a hot, sunlike star, which makes transits more likely to be spotted (because the orbit of a close-in planet does not have to be so precisely aligned for it to pass in front of the star). Finally, there are vastly more M dwarfs in the Milky Way than there are sunlike stars—about 250 of the former lie within 30 or so light-years of Earth, compared with only 20 of the latter.

GJ 1214b is not quite a second Earth: it is 2.7 times wider and six and a half times as massive as Earth, which gives it an overall density in between that of Earth and Neptune. Unfortunately, as Charbonneau and others realized immediately after the planet was discovered, this density can come about in several different ways. GJ 1214b could, for example, have a small, rocky core surrounded by a huge atmosphere of mostly hydrogen. It could also have a bigger core surrounded by a deep ocean of water, with a thin, water-rich atmosphere on top. It is impossible, given the density alone, to distinguish between those two possibilities—although the possibility of an ocean world is naturally more exciting, given that liquid water is considered a prerequisite for, if not a guarantee of, life as we know it.

Yet when University of Chicago astronomer Jacob Bean observed the planet in various wavelengths, hoping to see a change in its apparent size that would indicate just how thick the atmosphere is, he saw nothing. This could mean one of two things. The planet could have a puffy hydrogen atmosphere but one full of clouds and haze that would make it hard to detect. Or it could have a thin, watery atmosphere but one too thin to delineate with ground-based telescopes. The situation could be analogous to looking at a mountain range from a distance, says Kreidberg, who began working with Bean in 2013. "There may be peaks," she explains, "but if you're too far away, they might look like a flat line."

To try to resolve the issue, Bean and his colleagues were awarded 60 orbits of the Hubble. This was not the first time astronomers observed GJ 1214b with the Hubble, but it is by far the most intensive program to date. Alas, when the team analyzed their new data, they could neither confirm nor refute the possibility that GJ 1214b is a waterworld. Instead the Hubble observations only bolstered the case that, whatever GJ 1214b's composition might be, its skies are filled with obscuring clouds.

THE HUNT FOR OXYGEN

NOW THAT ASTRONOMERS have been in the planet-hunting business for some time, they have begun to find many more planets with long orbital periods. These planets are farther away from their stars and thus cooler than the early population of hot Jupiters. "For a long time we were limited to things that were 1,500 kelvins, 2,000 kelvins, so really quite hot," Knutson says. In these conditions, "most of the carbon in the atmosphere gets bound up with oxygen, forming carbon monoxide," she says. "The really interesting thing that happens as you drop below around 1,000 kelvins is that it switches to being incorporated into methane instead."

Methane is especially intriguing because it could be a sign of biological activity—though an ambiguous one because methane can be produced through purely geophysical processes. Oxygen—and especially ozone, a highly reactive molecule made from three oxygen atoms—would be far more likely to signal the presence of life. It would also be extremely difficult to detect because its spectral signature is subtle, especially so in the relatively small atmosphere of an Earth-size planet.

Yet for all the activity around medium-hot super-Earths, astronomers are still focused on the grand prize. "All of this is really just an exercise," Seager says. "I mean, it's interesting in and of itself, but for people like me, it's just a stepping-stone to when we finally get from super-Earths to studying the atmospheres of Earths."

That likely won't happen before the James Webb Space Telescope is launched into orbit, probably in 2018, and a new generation of huge, ground-based instruments, including the Giant Magellan Telescope and the Thirty Meter Telescope, come online in the 2020s. Even with those powerful instruments, Seager says, "it's going to take hundreds and hundreds of hours" of observing time. It is not clear even then that it will be possible to detect the signature of life unambiguously; for that, observers might still need the Terrestrial Planet Finder, whose funding has been reduced so drastically that any hope of an actual launch date is pure guesswork at this point.

Yet it is remarkable that, so far ahead of any schedule anyone dreamed of in the 1990s, Seager can even talk about the realistic prospect of finding biosignatures. We are no longer merely hoping that an alien civilization will spot us and point a message our way. We are actively exploring the air above distant worlds, searching their skies for signs that something is home. SA

Michael D. Lemonick is a writer at Climate Central, a nonprofit news site, and author of *Mirror Earth: The Search for Our Planet's Twin* (Walker Books, 2012). For 21 years he was a science writer for *Time* magazine.

MORE TO EXPLORE

Planets We Could Call Home. Dimitar D. Sasselov and Diana Valencia in *Scientific American*, Vol. 303, No. 2, pages 38–45; August 2010.

Exoplanet Atmospheres. Sara Seager and Drake Deming in *Annual Review of Astronomy and Astrophysics*, Vol. 48, pages 631–672; September 2010.

The Kepler exoplanet-detection mission: **http://kepler.nasa.gov**

scientificamerican.com/magazine/sa

© 2014 Scientific American

© 2014 Scientific American

STELLAR DYNAMICS

THE LONG-LOST SIBLINGS OF THE SUN

The sun was born in a family of stars. What became of them?

By Simon F. Portegies Zwart

PEOPLE HAVE OFTEN SOUGHT SOLITUDE in the starry night sky, and it is an appropriate place for that. The night is dark because, in cosmic terms, our sun and its family of planets are very lonely. Neighboring stars are so far away that they look like mere specks of light, and more distant stars blur together into a feeble glow. Our fastest space probes will take tens of thousands of years to cross the distance to the nearest star. Space isolates us like an ocean around a tiny island.

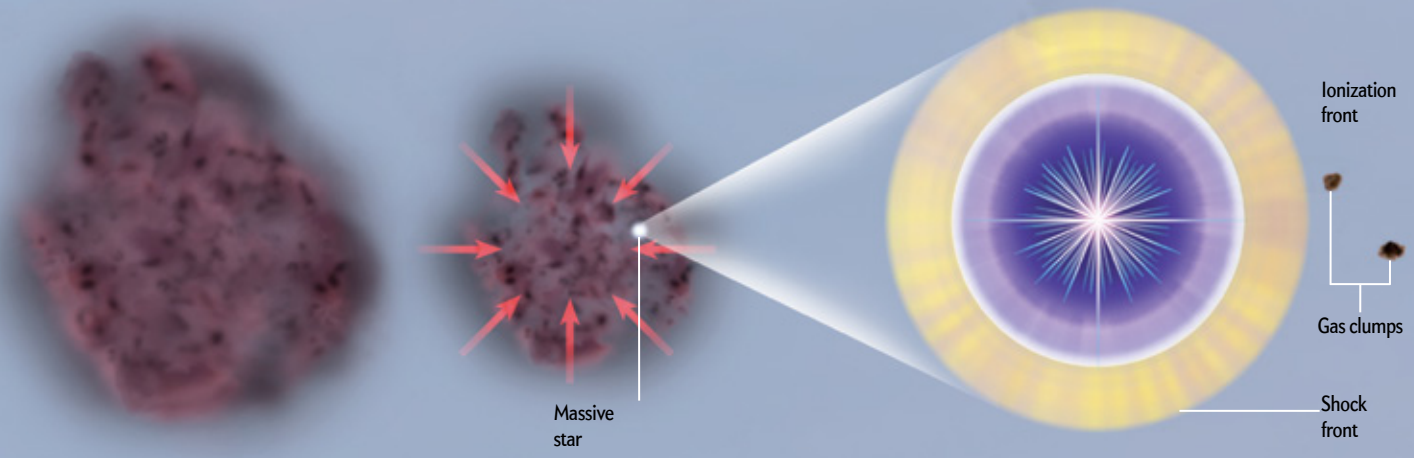
HAD YOU BEEN ALIVE at the dawn of the solar system, the night sky would have been bright enough to read by. A thousand or so stars formed within a few light-years from the same interstellar cloud the sun did.

Illustration by Ron Miller

© 2014 Scientific American

The Birth of the Sun's Cluster

Based on observations of star clusters and the inferred properties of the cluster into which the sun was born, J. Jeff Hester and Steven J. Desch of Arizona State University and their colleagues have reconstructed the events leading up to the formation of the sun.



A giant cloud of molecular gas accumulates and begins to collapse under its own weight.

One or more massive stars form in the densest regions of the cloud.

Each massive star pours out ultraviolet radiation, ionizing the surrounding gas and driving out a shock front. The shock expands at a few kilometers per second.

Yet not all stars are so secluded. About one in 10 belongs to a cluster, a swarm of hundreds to tens of thousands of stars with a diameter of a few light-years. In fact, most stars are born in such groups, which generally disperse over billions of years, their stars blending in with the rest of the galaxy. What about our sun? Might it, too, have come into existence in a star cluster? If so, our location in the galaxy was not always so desolate. It only became so as the cluster dispersed in due time.

A growing body of evidence suggests just that. Although conventional wisdom once held that the sun was an only child, many astronomers now think it was one of about 1,000 siblings all born at nearly the same time. Had we been around at the dawn of the solar system, space would not have seemed nearly so empty. The night sky would have been filled with bright stars, several at least as bright as the full moon. Some would have been visible even by day. Looking up would have hurt our eyes.

The cluster into which the sun was probably born is now long gone. I have pieced together the available data and made an educated guess as to what it might have looked like. From these inferred properties, I have calculated the possible trajectories of former cluster members through the galaxy to figure out where they might have ended up. Although they have scattered and intermixed with millions of unrelated stars, they should be identifiable with the European Space Agency's Global Astrometric Interferometer for Astrophysics (GAIA) satellite, which was launched in 2013. The motions of their orbits and their sunlike chemical compositions should give them away. And reuniting with our long-lost stellar siblings should enable astronomers to reconstruct the conditions under which a shapeless cloud of gas and dust coalesced into our solar system.

MEMORIES OF OUR BIRTH

THE MOST COMPELLING EVIDENCE that the sun has siblings emerged in 2003, when Shogo Tachibana, now at Hokkaido University, and Gary R. Huss, now at the University of Hawaii at Manoa, analyzed two primitive meteorites that are thought

IN BRIEF

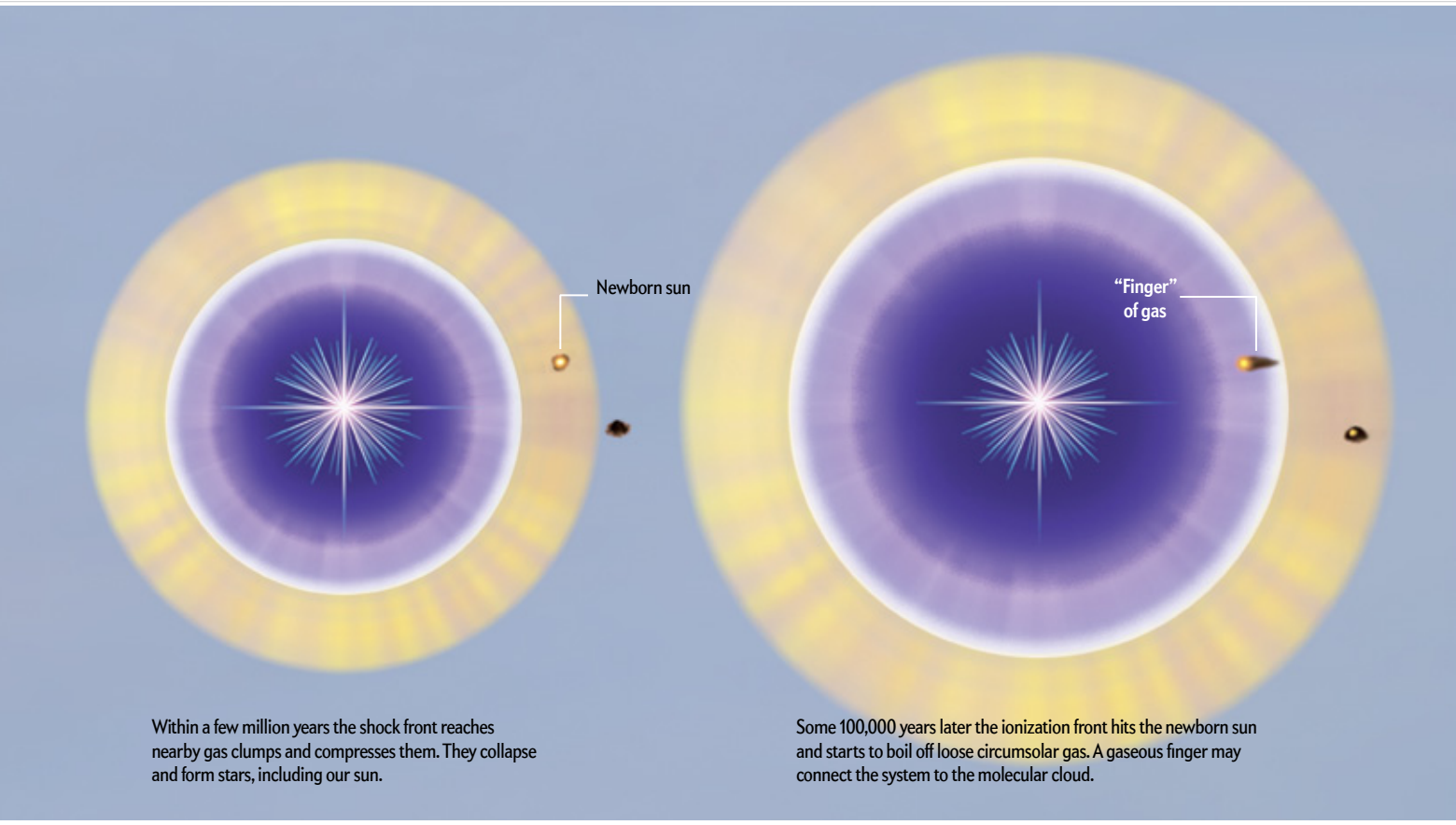
The sun is a solitary star, and astronomers have traditionally assumed it formed as such. Yet most stars are born in clusters, and scraps of evidence from meteorites and from the arrangement of comets suggest that our sun was no exception.

Its birth cluster could have contained 1,500 to 3,500 stars within a diameter of 10 light-years—a big, unhappy family whose larger members bullied the small fry and which broke up not long after our solar system came into being.

Although the sun's siblings have long since dispersed across the galaxy, observatories such as the European GAIA satellite will be able to look for them. Their properties might fill in the gaps of the solar system's deep history.

Illustrations by Don Dixon

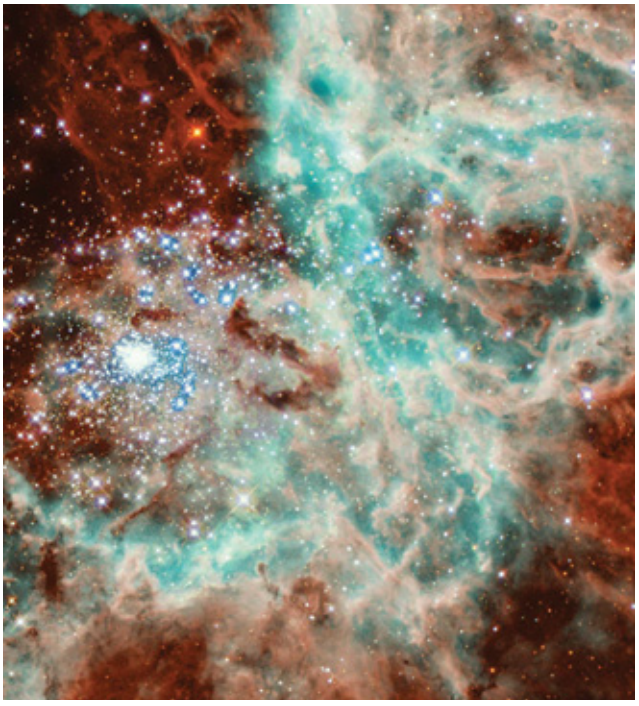
© 2014 Scientific American


Newborn sun
"Finger" of gas
Within a few million years the shock front reaches nearby gas clumps and compresses them. They collapse and form stars, including our sun.
Some 100,000 years later the ionization front hits the newborn sun and starts to boil off loose circumsolar gas. A gaseous finger may connect the system to the molecular cloud.

to be almost pristine leftovers of solar system formation. They detected nickel 60, the product of the radioactive decay of iron 60, in chemical compounds where, by rights, iron should be found. It seems a game of chemical bait and switch took place in the meteorite: the compounds originally formed from iron, the iron metamorphosed into nickel, and the nickel was locked in place, forever an interloper.

The iron 60 had to be synthesized and injected into the solar system's meteorites within its radioactive half-life, which recent estimates peg at approximately 2.6 million years. That is a cosmic eyeblink. Therefore, the iron had to come from very nearby—and the likeliest source is a supernova explosion. Based on various isotopic measurements, Leslie Looney of the University of Illinois and his co-workers argued in 2006 that a supernova must have detonated within a distance of five light-years when the sun was only 1.8 million years old. The supernova might have been as close as 0.07 light-year.

If the sun had been as secluded as it is today, the location and timing of the supernova would be quite a coincidence. Was a massive star simply passing by when it decided to blow up? No other supernova has ever gone off at such close range; if it had, it would probably have wiped out life on Earth. A much more plausible explanation is that the newborn sun and the exploding star were fellow members of a cluster. With stars packed so tightly together, a close supernova would not have been so improbable.



STAR CLUSTER R136, located in a region known as the Tarantula nebula, is similar to (but much denser than) the cluster into which the sun was born.

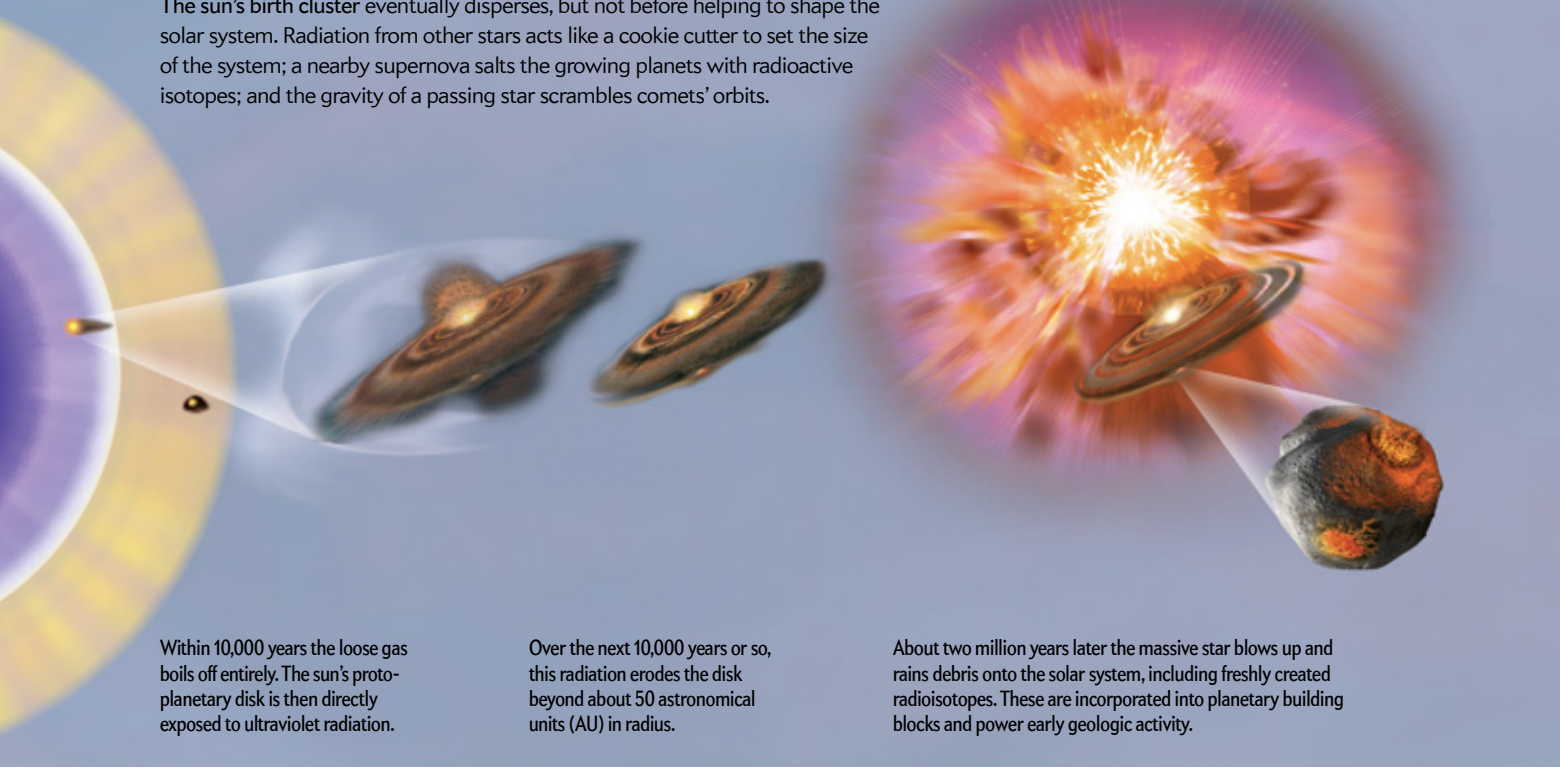
NASA, N. WALBORN AND J. MAÍZ-APELLÁNIZ (STScI) AND R. BARBÁ (*La Plata Observatory, Argentina*)

© 2014 Scientific American

TIMELINE CONTINUED

The Death of the Sun's Cluster

The sun's birth cluster eventually disperses, but not before helping to shape the solar system. Radiation from other stars acts like a cookie cutter to set the size of the system; a nearby supernova salts the growing planets with radioactive isotopes; and the gravity of a passing star scrambles comets' orbits.



Within 10,000 years the loose gas boils off entirely. The sun's proto-planetary disk is then directly exposed to ultraviolet radiation.

Over the next 10,000 years or so, this radiation erodes the disk beyond about 50 astronomical units (AU) in radius.

About two million years later the massive star blows up and rains debris onto the solar system, including freshly created radioisotopes. These are incorporated into planetary building blocks and power early geologic activity.

THE STARRY CLUSTERS BRIGHT

THE IDEA THAT THE SUN originated in a star cluster is at odds with the classical view of clusters that is still common in textbooks. Astronomers have traditionally classified clusters into two types: so-called galactic, or open, clusters and globular clusters. The former are young, sparsely populated and located primarily in or near the plane of our galaxy. The prototypical example is Praesepe, also known as the Beehive cluster or as M44. It was one of the first objects at which Galileo pointed his telescope 400 years ago, in 1609. What looked like a splotch of light revealed itself as an array of stars—up to 350 of them, all born about 700 million years ago.

In contrast, globular clusters are very old, densely populated and located all around the galaxy, not just in a plane. The first was discovered in 1746 by Italian astronomer Giovanni Maraldi and is now known as M15. It contains about a million stars with an age of about 12 billion years.

The trouble is that neither category fits the sun. Its advanced age of 4.6 billion years suggests it should have been born in a globular cluster, yet its location points to a galactic cluster. Recently, however, we have realized that not all clusters fall neatly into one of these two classical types [see "The Unexpected Youth of Globular Clusters," by Stephen E. Zepf and Keith M. Ashman; SCIENTIFIC AMERICAN, October 2003].

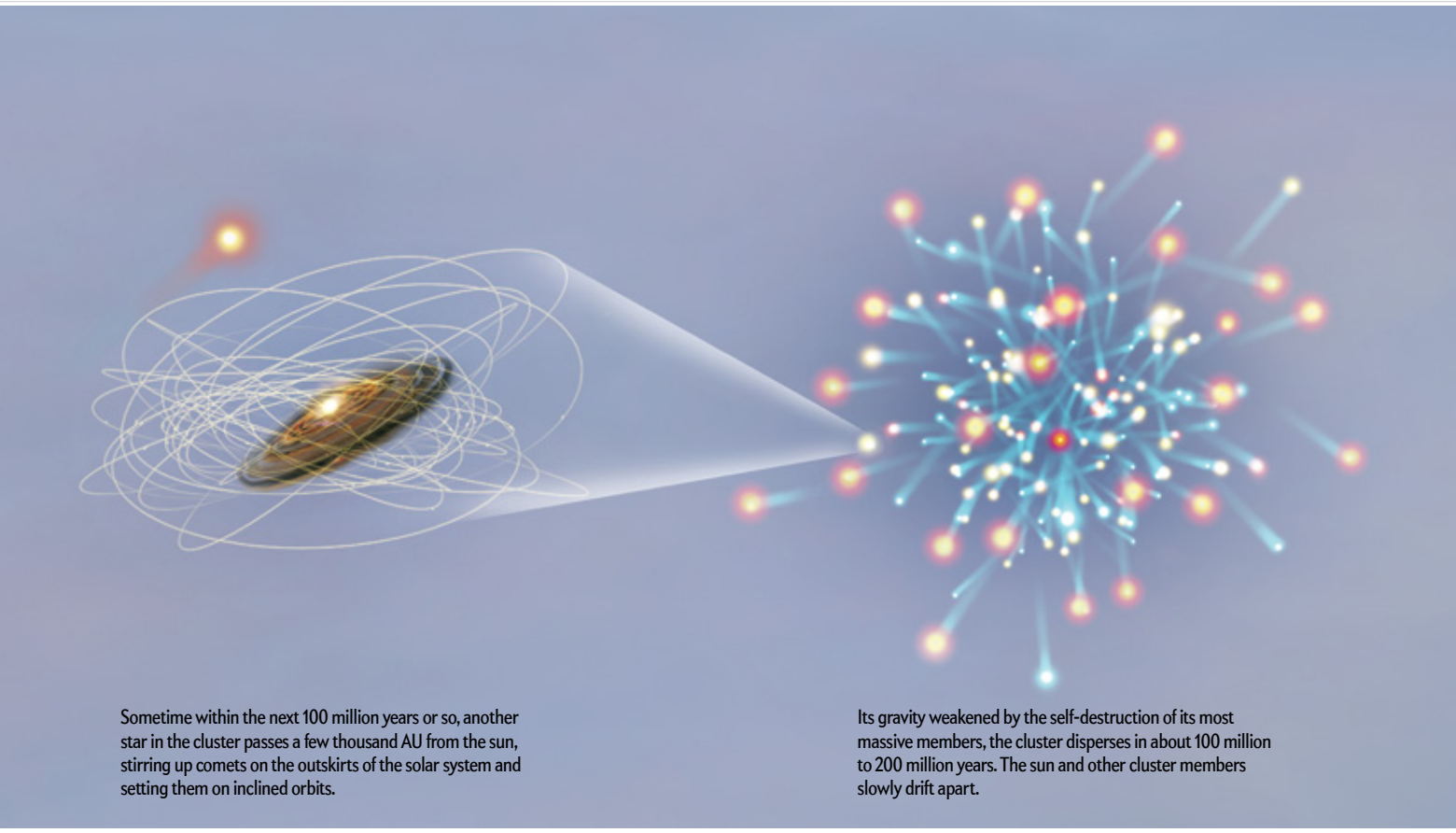
What changed our minds was the star cluster R136, which is located in one of the Milky Way's small satellite galaxies, the Large Magellanic Cloud. First spotted in 1960, R136 was initially thought to be a single, giant star 2,000 times as massive as the sun and 100 million times as bright. But in 1985 Gerd Weigelt and Gerhard Baier, both then at the University of Erlangen-Nürnberg in Germany, used new high-resolution imaging techniques to show that R136 is actually a cluster of about 10,000 stars a few million years old. It is as dense as a globular but as young as a galactic cluster. With characteristics of both types, R136 was the missing link between them. Since then, observers have found several clusters like R136 in our galaxy. Other galaxies such as the Antennae contain hundreds if not thousands of them.

The discovery that stars continue to form in clusters so dense they could be mistaken for a single star was astonishing. It led to considerable consternation among theorists. On the one hand, we were relieved because we had not been able to explain R136 as a single superstar. On the other hand, we had to reconsider everything we thought we knew about star clusters. We now think that the majority of stars, including the sun, are born in tight clusters such as R136. A cluster forms out of a single gas cloud and, over time, evolves into either a galactic or globular cluster depending on its mass and environment.

DREAMS FROM OUR STELLAR FATHERS

THE MEMBERS OF A CLUSTER span a range of masses, with a few heavy stars and a multitude of lightweight ones. The least massive, with a tenth the mass of the sun, are the most common, and for every factor of 10 increase in mass, the abundance of stars drops by about a factor of 20.

© 2014 Scientific American



Sometime within the next 100 million years or so, another star in the cluster passes a few thousand AU from the sun, stirring up comets on the outskirts of the solar system and setting them on inclined orbits.

Its gravity weakened by the self-destruction of its most massive members, the cluster disperses in about 100 million to 200 million years. The sun and other cluster members slowly drift apart.

Thus, for each star of 15 to 25 solar masses—the size of the one that went supernova near the newborn sun—a cluster contains some 1,500 lesser stars. This number sets the minimum mass of the sun's birth cluster. The maximum mass is set by the fact that the larger a cluster is, the longer it takes for massive stars to settle toward the center, where they have the greatest likelihood of affecting their smaller brethren. Based on my simulations, the cluster probably contained fewer than about 3,500 stars.

A star of 15 to 25 solar masses lives for six million to 12 million years before blowing up, so it must have formed about this long before the sun did. In other clusters, such as the famous Trapezium cluster in the Orion nebula, astronomers have found that massive stars are usually the first to form, with sunlike stars arising several million years later.

A cluster of the inferred mass was too flimsy to evolve into a globular cluster. Instead it dispersed after 100 million to 200 million years. The massive stars at its center shed gas in stellar winds (similar to but much more intense than the solar wind) and eventually exploded, reducing the density of material in the cluster and thereby weakening its gravitational field. Consequently, the cluster expanded and might have fallen apart. Even if it survived this early outgassing, interactions among stars and the tidal forces exerted by the rest of the galaxy drove its slow dissolution.

Before the cluster disintegrated, stars were so densely packed that one could easily have passed through the solar system. A stellar close encounter would have pulled planets, comets and asteroids from their original circular, planar orbits into highly elliptical and inclined orbits. Many comets beyond a distance of 50 astronomical units (AU), past the orbit of Neptune, have highly skewed orbits. The internal dynamics of the solar system seem incapable of accounting for these peculiar orbits; the bodies are beyond the gravitational influence even of Jupiter. The

EVIDENCE

Not an Only Child

Several telltale signals suggest that the sun was born in a cluster:

Ancient meteorites contain the decay products of short-lived radionuclides such as iron 60 and aluminum 26. The source of the isotopes (probably a supernova) must have been very nearby, indicating that the early sun was not alone.

The sun's levels of heavy elements are higher than its location in the galaxy would otherwise indicate—suggesting that it was topped up with debris from a nearby supernova.

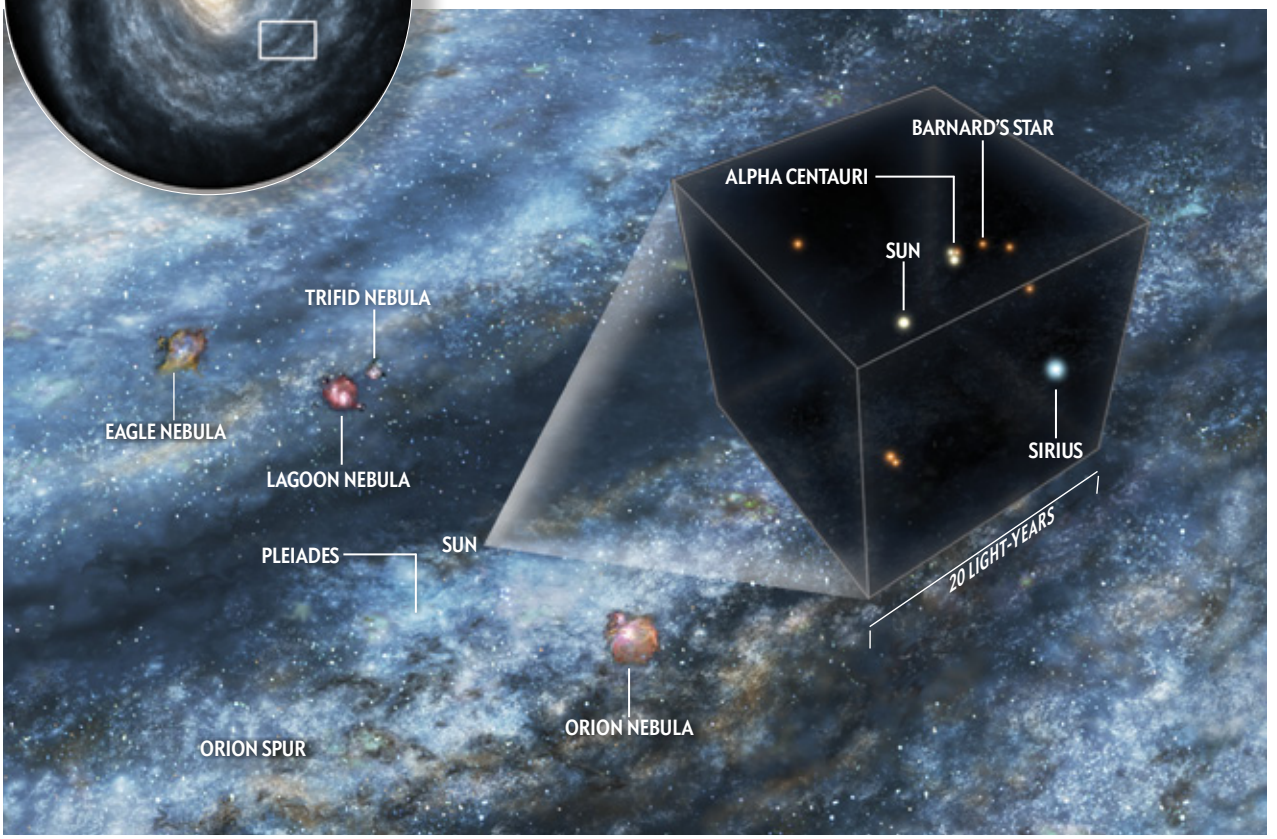
Uranus and Neptune are much smaller than Jupiter and Saturn. One reason might be that the radiation of a nearby star boiled off their outer layers. Planets closer to the sun avoided this fate because residual interplanetary gas shielded them.

© 2014 Scientific American

GEOGRAPHY OF THE GALAXY

Meet the Neighbors

The sun currently lies about 30,000 light-years from the center of the Milky Way galaxy. Astronomers know of only about a dozen other stars within 10 light-years of the sun. Before the sun's cluster dispersed, the same volume held more than 1,000 stars.



most likely explanation is that they were stirred up by a star passing 1,000 AU away. The planets, though, have very regular orbits, indicating that no stellar intruder ever came within 100 AU of the sun.

From these facts, I have estimated the dimensions of the cluster. For another star in the cluster to pass 1,000 AU away with reasonable probability over the cluster lifetime, the cluster had to be less then 10 light-years in diameter. Conversely, for a star not to come within 100 AU, the cluster had to be greater than three light-years in diameter. In short, the sun's birth cluster looked like R136 but much less dense, so that stars were far enough apart not to interfere with planet formation.

SOLAR GENEALOGY

THEORISTS CAN GO FURTHER and ask where exactly in the galaxy the birth cluster was located. The solar system revolves around the galactic center in an almost circular orbit, more or less in the disk. At the moment, we are located about 30,000 light-years from the center and about 15 light-years above the plane of the disk, orbiting at a speed of 234 kilometers per second. At this rate, the sun has done 27 circuits since its formation. Its orbit is not a closed loop but a somewhat more complicated shape determined by the gravitational field of the galaxy, which astronomers infer from the motion of stars and interstellar gas clouds.

Assuming, provisionally, that the gravitational field has not changed over the past 4.6 billion years, I have projected the orbit backward in time and deduced that the sun was born 33,000 light-years from the center and 200 light-years above the galactic plane. What makes this position puzzling is that the outer reaches of the galaxy are poorer in heavy elements than the inner parts. The most distant regions may lack enough material to make planets, let alone life [see "Refuges for Life in a Hostile Universe," by Guillermo Gonzalez, Donald Brownlee and Peter D. Ward; SCIENTIFIC AMERICAN, October 2001]. Although the sun's putative birthplace is not quite so impoverished, it is still poorer in heavy elements than the sun is. Based purely on the sun's heavy element composition, astronomers would have expected it to form 9,000 light-years closer to the center.

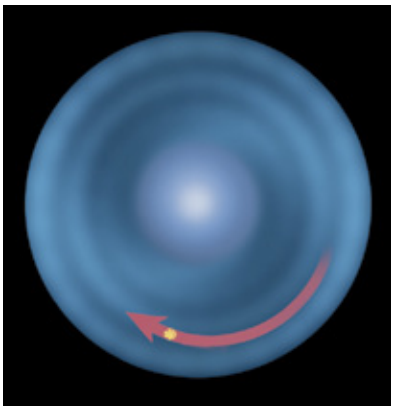
Maybe the supernova that seeded meteorites with iron 60 also enriched the sun with heavy elements. Or maybe my orbit-

© 2014 Scientific American

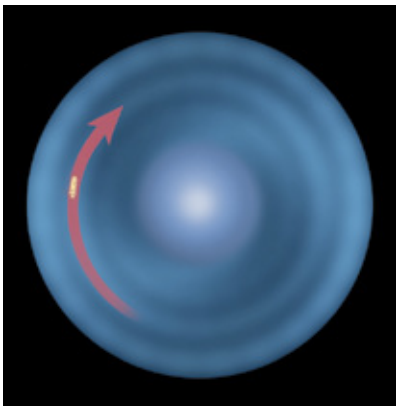
CLUSTER DISPERSAL

Family Breakup

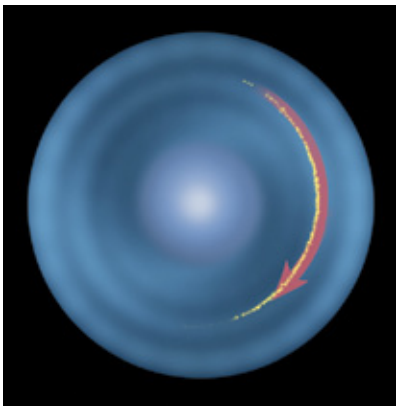
By tracking the likely orbits of the sun's siblings as they dispersed, theorists can estimate where they ended up, so that observers can search for them. The discovery of even one would help reconstruct the origin of the solar system.



The sun and its siblings move apart at a few kilometers per second, while continuing to orbit the galaxy's center at more than 200 kilometers per second.



After doing a fourth of an orbit around the galactic center, the stars have spread out only 100 or so light-years.



After making 27 full orbits, the stars form a stream tens of thousands of light-years long. A few percent of them still lie within 300 light-years of the sun.

al calculation went astray because the gravitational field of the galaxy has changed or because the sun's orbital path was diverted slightly by the gravity of nearby stars or gas clouds. In that case, the sun was born closer to the center than I estimated, and its composition is not so anomalous.

The sun's ex-family members, too, should be orbiting around the galactic center at more than 200 kilometers per second. Yet their relative velocity, which is determined by their mutual gravitational forces in the original cluster, is only a few kilometers per second. Like clumps of cars on a highway, they stick together even though they are no longer bound to one another gravitationally. The original swarm has spread into an arc only very gradually. After 27 orbits, it should stretch about halfway around the galaxy.

My calculations suggest that about 50 of the sun's brothers and sisters should still be within 300 light-years of our current location and that about 400 stars are within 3,000 light-years. Depending on the stars' original relative velocity and the timing of their departure from the cluster, the sun either follows in their orbital footsteps or they in ours.

The best place to look for them is in the plane of the galaxy in the direction the solar system is moving or in exactly the opposite direction. We have been looking for them in a catalogue of stars assembled by ESA's Hipparcos satellite [see "The Star Mapper," by Philip Morrison; Scientific American, February 1998]. But the Hipparcos data are not precise enough to identify any siblings. A breakthrough came earlier in 2014, when Ivan Ramirez of the University of Texas at Austin and his co-workers measured the chemical composition of the yellow dwarf star HD162826. They found it to be very similar to our sun's, as if it had been exposed to the same supernova that enriched our solar system. HD162826 could be one of our lost siblings, although more detailed information will provide more proof. Such proof could come from the ESA satellite GAIA, which has two telescopes measuring the position and velocity of some one billion stars over five years, creating an essentially complete census of stars brighter than 22nd magnitude within some 30,000 light-years of the sun. The solar siblings will stand out in these data because they lie nearly along the sun's past and future path. Identifying even one sibling of the sun, or confirming HD162826's genealogy, will provide crucial information about the very early days of the solar system. Theorists will be able to compute the birthplace of the sun with greater certainty and determine, for example, whether the gravitational field of the galaxy has changed substantially or not. Solar siblings will also be excellent places to look for habitable planets. Although we seem very alone in the galaxy, it was not always this way. Many of the sun's seeming idiosyncrasies—particularly its nurturing of life—might make sense in the context of its family. SA

Simon F. Portegies Zwart is a professor of computational astrophysics at Leiden Observatory of Leiden University in the Netherlands. His specialties are high-performance computing and gravitational stellar dynamics, in particular the ecology of dense stellar systems. In his spare time, he translates Egyptian hieroglyphs and brews beer.

MORE TO EXPLORE

The Formation of Star Clusters. Bruce Elmegreen and Yuri Efremov in *American Scientist*, Vol. 86, No. 3, pages 264–273; May-June 1998.

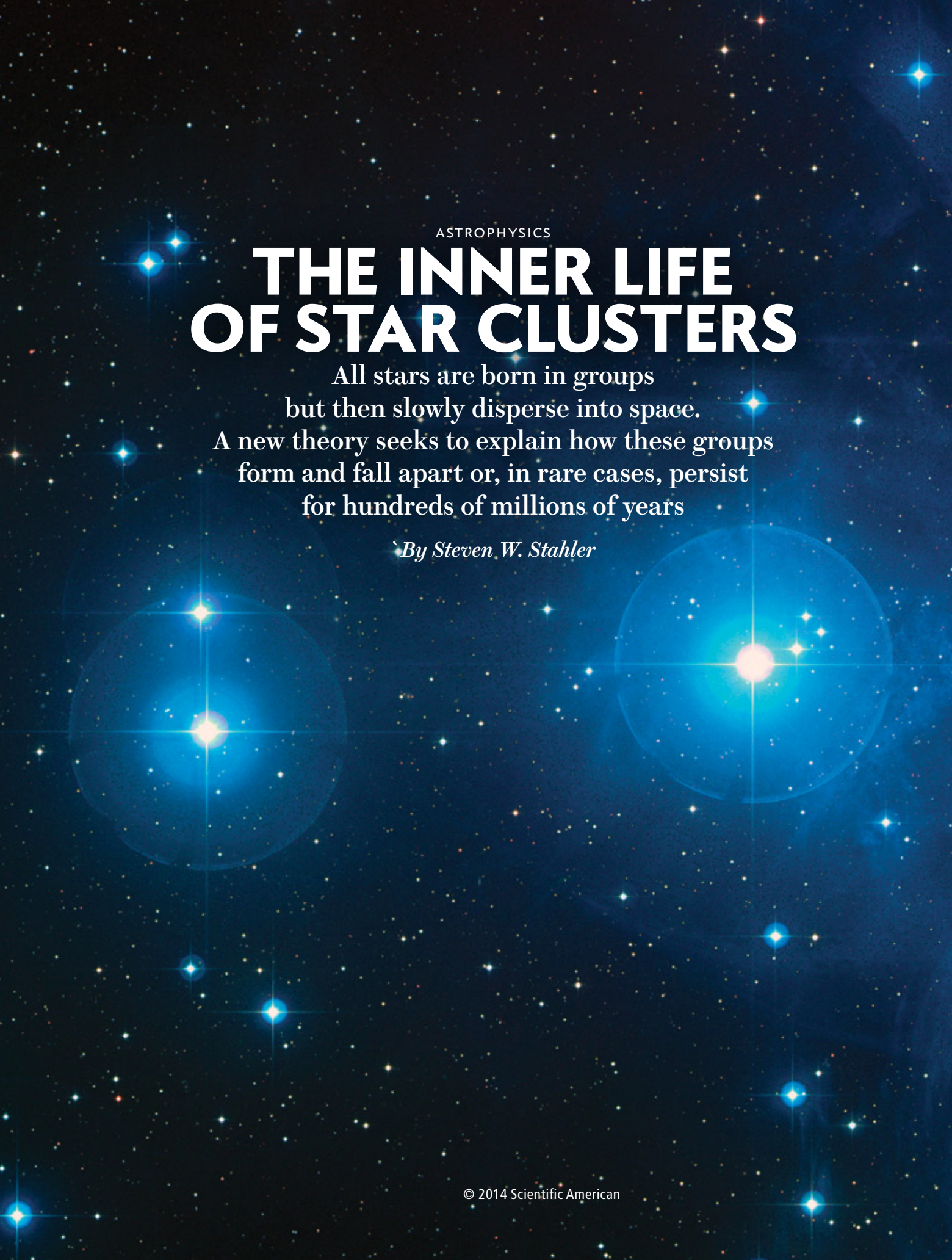
The Cradle of the Solar System. J. Jeff Hester, Steven J. Desch, Kevin R. Healy and Laurie A. Leshin in *Science*, Vol. 304, pages 1116–1117; May 21, 2004.

Radioactive Probes of the Supernova-Contaminated Solar Nebula: Evidence That the Sun Was Born in a Cluster. Leslie W. Looney, John J. Tobin and Brian D. Fields in *Astrophysical Journal*, Vol. 652, No. 2, pages 1755–1762; December 1, 2006. Available online at arxiv.org/abs/astro-ph/0608411

The Lost Siblings of the Sun. Simon F. Portegies Zwart in *Astrophysical Journal Letters*, Vol. 696, No. 1, pages L13–L16; May 1, 2009. arxiv.org/abs/0903.0237

scientificamerican.com/magazine/sa

© 2014 Scientific American



ASTROPHYSICS

THE INNER LIFE OF STAR CLUSTERS

All stars are born in groups but then slowly disperse into space. A new theory seeks to explain how these groups form and fall apart or, in rare cases, persist for hundreds of millions of years

By Steven W. Stahler

© 2014 Scientific American

STAR STUCK: The Pleiades, known as an "open" cluster, is one of the most stable stellar groups in the Milky Way.

© 2014 Scientific American

THE NIGHT SKY IS A FIELD OF STARS. IN EVERY DIRECTION, STARS BRIGHT and dim fill the horizon to brimming. Some seem to form distinct patterns, which we recognize as constellations. Yet as beguiling as those patterns may be, most of them are no more than projections of the human mind. The vast majority of stars, in our own galaxy and in others, have no true physical connection to one another. At least, not anymore. Every star actually begins its life in a group, surrounded by siblings of nearly the same age that only later drift apart. Astronomers know this because some of these stellar nurseries, called star clusters, still exist. The Orion nebula cluster is perhaps the most famous one: in images from the Hubble Space Telescope, its stars wink from within churning clouds of dust and gas. You can see the Pleiades cluster from your backyard: it is the fuzzy patch in the constellation Taurus.

Star clusters vary enormously, ranging from fragile associations with just a few dozen members to dense aggregates of up to a million stars. Some groups are very young—only a few million years old—and others date from the dawn of the universe. Within them, we find stars in every stage of the stellar life cycle. Indeed, observations of star clusters provided the main evidence for today's accepted theory of how individual stars evolve over time. The theory of stellar evolution is one of the triumphs of 20th-century astrophysics.

Yet relatively little is known about the inner workings and evolution of the clusters themselves. What accounts for the variety of forms that astronomers observe? We understand far more about individual stars than we do about the cradles of their formation!

The irony of this situation first struck me 20 years ago, when I began writing a graduate textbook on star formation with Francesco Palla of Arcetri Astrophysical Observatory in Florence, Italy. At the time, the two of us were regularly trading visits between Berkeley, Calif., and Florence. As we followed the many strands of research in this rich field, the unanswered questions about star clusters always lurked in the back of our minds.

One afternoon, as we took a break at Caffe Strada (located in Berkeley, naturally), the germ of an answer came to me. Perhaps the same physical forces had shaped all clusters, regardless of their present ages and sizes. And perhaps one simple variable could account for the way those forces act on an individual cluster: the mass of the parent cloud from which each cluster is born. It would take me the better part of the ensuing decades to gather evidence for this hunch.

CLOUDY WITH A CHANCE OF STARLIGHT

WHEN I BEGAN THIS WORK, astronomers knew a lot about how stars form and a good bit about the kinds of clusters they form in. Stars do not materialize from empty space; rather they coalesce within vast clouds composed chiefly of hydrogen molecules, along with other elements and a small admixture of dust. These so-called molecular clouds are distributed throughout all galaxies, and each exerts a gravitational pull—not only on stars and other objects outside the cloud but also on regions in the cloud itself. Because of the cloud's own gravity, regions where gas and dust are especially dense collapse into protostars. In this way, clusters of anywhere from a few dozen to thousands of stars can arise from a single molecular cloud.

Clusters generally occur in five types, distinguished in part by their age and in part by the number and density of the stars they contain. The very youngest stellar groups, called embedded clusters, lie in clouds so thick that light in the visible wavelengths radiating from their member stars is completely obscured. Instead we see only the infrared glow of dust heated by the embedded stars and cannot discern the detailed structure

IN BRIEF

Stars form in clusters, within clouds composed of gas mixed with dust.
Three types of clusters can be seen in the Milky Way, with differing structures and evolutionary histories.
The mass of the cloud that spawns a stellar group may account for these differences, by affecting the balance of contraction and expansion in the cluster.
Only open clusters remain intact after the parent cloud has dispersed.

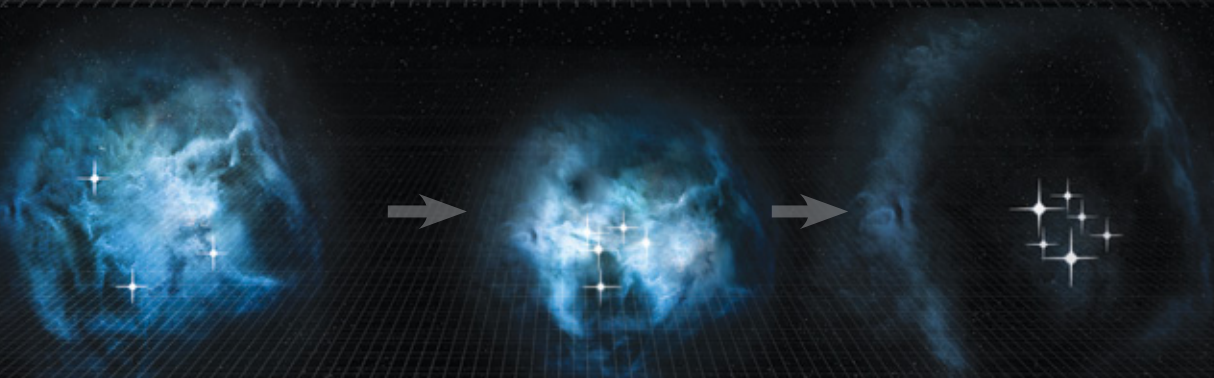
PRECEDING PAGES: COURTESY OF NASA, ESA, AURA AND CALTECH

© 2014 Scientific American

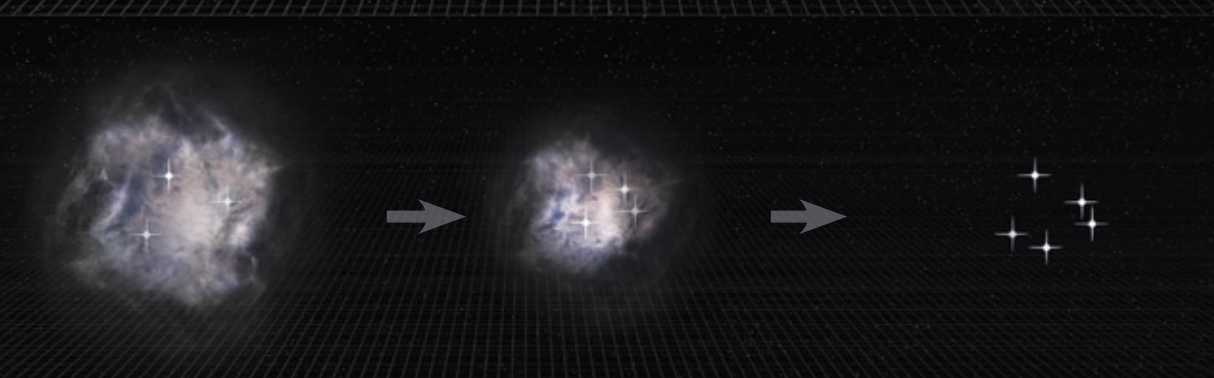
One Driver, Three Outcomes

The three most readily observed star clusters in our Milky Way galaxy all began as a diffuse cloud of dust and gas, within which small regions condensed to form stars. The author proposes that a single factor—the mass of the parent cloud—accounts for the differences in the clusters' subsequent evolution and structure (*below*). First, the clouds contracted, accelerating star production at rates determined by the starting mass; higher-mass clouds contracted most and generated stars most quickly. Later, the clusters expanded, and the clouds dispersed partly or fully, depending on the number and kinds of stars present.

T Association Such clusters typically last just a few million years and possess up to several hundred young stars, called T Tauri, surrounded by remnants of the parent cloud. They probably arise when a low-mass cloud contracts slowly under the influence of its own relatively weak gravitational pull. The mild contraction would produce a smattering of stars. Stellar winds from these stars would later diminish the cloud, causing both gas and stars to spread apart.



OB Association These clusters stay bound together for as long as 10 million years and comprise thousands of densely packed stars, including a few very massive ones called O and B stars. To produce such a dense cluster, the parent cloud would have to be extremely massive and contract very quickly. Harsh ultraviolet radiation from the most massive stars would then shred the parent cloud, and the cluster would expand and eventually disperse.



Open Cluster The longest lasting of the three, open clusters can persist for hundreds of millions of years, but they have far fewer stars than OB associations do. These groups probably arise when a cloud of intermediate mass contracts. Although stellar winds drive the cloud away, the cluster does not disperse for a very long time.

Illustrations by Malcolm Godwin

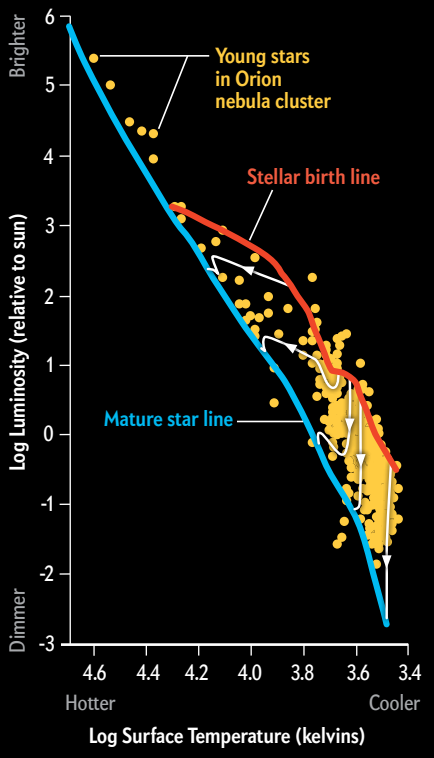
© 2014 Scientific American

Cluster's History Supports Theory

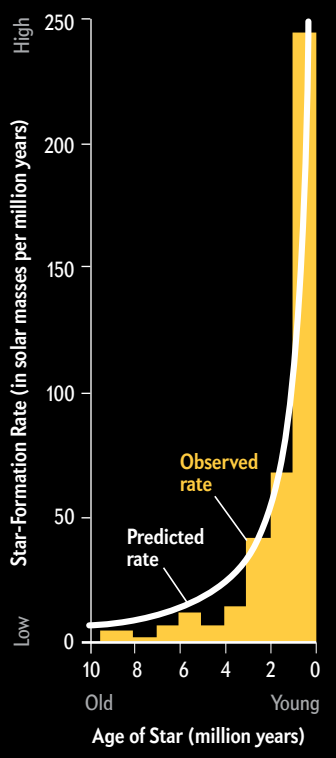
Data from the Orion nebula cluster, an OB association in the Orion nebula (*photograph*), support the author's theory that cloud contraction occurs early in cluster evolution, causing star formation to accelerate as the density of the parent cloud rises. Star formation in the cluster stopped about 100,000 years ago, but until then the parent cloud probably contracted for millions of years. To show that contraction happened, the author first determined the ages of the cluster's young stars (*yellow dots in left graph*), which can be gleaned from their temperature and brightness. In general, the red line represents very young stars that are newly visible in optical wavelengths; these stars get hotter and dimmer in predictable ways as they age (*white arrows in left graph*), until ultimately they are plotted on the blue line. Thus, a star's position between the red and blue lines indicates its age.

The author then calculated the collective mass of the stars (relative to the mass of the sun) in each million-year age group, revealing the star-formation rate in the cluster during that period (*right graph*). The results indicate that star production increased dramatically over time, just as theoretical modeling (*white line*) predicted.

Young Stars Show Their Age



Star Formation Has Accelerated



SOURCES: "STAR FORMATION IN THE ORION NEBULA CLUSTER" BY FRANCESCO PALLA AND STEVEN W. STAHLER, IN *ASTROPHYSICAL JOURNAL*, VOL. 525; 1999 (*left*); COURTESY OF NASA, HST, C. R. O'DELL AND S. K. WONG (*right*)

of these primitive clusters. They remain an abiding mystery.

Globular clusters, in contrast, are the oldest and most populous stellar group. They date to the dawn of the universe and can contain as many as one million stars packed very closely together. The parent clouds of these mature clusters have disappeared, and the stars within are entirely visible. Yet the closest globular clusters lie at some remove from the disk of the Milky Way galaxy, and so they, too, are difficult for astronomers to study in detail.

For practical reasons, then, I have restricted my theorizing to the three types of clusters that occur in the plane of our galaxy and thus can be seen best of all. The sparsest of these is called a T association because it consists largely of the most common kind of young star, called a T Tauri. (Our sun was a T Tauri star in its youth.) Each T association contains up to several hundred of these stars surrounded, but not totally obscured, by the parent cloud. T associations do not stay together for long: the most aged ones observed are some five million years old—the blink of an eye from a cosmic perspective.

Scientists have known for some time that the mass of the parent cloud in a T association is far greater than that of its collective stellar progeny. I believe this feature accounts for the short life span of these clusters. Mass determines the strength of gravitational force: the greater the mass is, the stronger the gravitational pull is. So if the mass of the parent cloud in a T association is much greater than that of its member stars, the gravity of the cloud—not the gravity exerted by the stars on one another—must be what holds the cluster together. And if the cloud disperses, the stars will drift apart. Astronomers think that stellar winds—jets of gas propelled forcefully outward from the stars—eventually strip away the parent cloud of a T association, freeing the previously bound stars to head into space.

© 2014 Scientific American

The second type of stellar group readily observed in the Milky Way is named after two extraordinary kinds of stars, designated O and B, that are the most luminous and the most massive stars in the universe. These clusters, called OB associations, typically have about 10 times more stars than T associations do, including a few O and B types. The Orion nebula cluster is a familiar example; located some 1,500 light-years away, it comprises four truly massive stars and about 2,000 lesser ones, including many T Tauri stars. It has the highest stellar density of any region in our part of the galaxy.

All young OB associations have similarly high densities and arise from especially massive parent clouds. Yet despite the tremendous gravity within these systems, the stars in older OB associations are not just gradually dispersing but are actively flinging themselves out into space. Astronomers know this because images of mature OB associations taken just a few decades apart show that the member stars have moved farther away from one another.

One reason for this rapid dispersal is that the stars are moving very quickly to begin with. The terrific gravity of the parent cloud in an OB association spurs its stellar members to orbit at high velocities. Young OB associations are chock-full of these speeding stars, poised to escape the cluster should the parent cloud diminish. And in OB associations, the parent cloud is under siege from the harsh ultraviolet radiation that O and B stars emit during their brief lifetimes. These stars are powered by nuclear fusion, as our sun is, but they burn far more fiercely. A typical O star has 30 times the mass of the sun, for example, yet it exhausts its fuel in just a few million years.

In the course of this self-immolation, ultraviolet radiation streams from the star and ionizes surrounding gas—in effect, burning up the parent cloud. The dust and gas in the Orion nebula cluster are glowing from this ionization. As the parent cloud burns away, its gravity dwindles. When the massive stars have finally expired and the parent cloud is gone, the gravity of the system can no longer contain those smaller, speeding stars, and they are flung far afield.

Thus, both T and OB associations undo themselves, whether through gentle attrition or violent agitation. The third, much less common, kind of stellar group in the Milky Way is remarkably stable, however. Called open clusters, these groups have up to 1,000 ordinary stars and persist for hundreds of millions, and even billions, of years. Yet their molecular clouds and any attendant gravity have long since vanished.

The Pleiades is one such cluster. It is 125 million years old, and its parent cloud has probably been gone for 120 million years or more. The equally famous Hyades cluster, not far from the Pleiades in the sky, is 630 million years old. In the outskirts of our galaxy dwell dozens of open clusters that are even older. The cluster M67, a system of 1,000 stars, arose four billion years ago.

Even open clusters are not immortal; very few are older than M67. Astronomers believe that, eventually, the gravity of molecular clouds passing close by begins to shred and disperse these systems. Yet they still pose a vexing problem. Over the past few decades researchers have arrived at satisfying explanations for how the dispersal of parent clouds causes T and OB associations to fall apart. But they still have no answer for why the stars in open clusters survive cloud dispersal to remain bound together for many millions of years.

PUSH AND PULL

AS I WROTE my book on star formation, I had ample cause to wonder about the diversity of cluster forms. I saw the mystery of open clusters as part of a larger class of questions: Why does our galaxy host only a limited variety of star clusters? How does a molecular cloud "decide" what kind of cluster it is going to produce?

I considered the forces at work in star clusters. Taken together, the life stages of the three types I chose to study point to two countervailing processes: contraction, caused by the gravity of the parent cloud, and expansion, promoted by stellar winds and ionizing radiation. Each star-producing cloud is subject to these two opposing influences to varying degrees. In the case of T and OB associations, expansion eventually wins. In the case of open clusters, expansion and contraction seem to stay in balance, at least during the critical epoch during which member stars are forming.

The balance of forces in a cloud, I reasoned, thus determines its fate as well as the destiny of the stellar cluster it produces. And I suspected that the key to this balance might be the original mass of the parent cloud. As I have explained, the mass of a cloud certainly determines its gravity; the cloud's gravity, in turn, governs the rate at which it contracts. Cloud mass also determines the number of stars the cloud produces. A low-mass cloud, for example, would contract slowly, causing a gradual increase in its density that would produce a smattering of ordinary stars. Later, winds from those stars would gradually strip away the cloud, reversing the contraction and releasing member stars into space. That scenario fits what we observe in T associations today.

At the opposite extreme, a cloud with an order of magnitude more mass would undergo a rapid contraction, forming many new stars in close proximity. Eventually this cloud's core would reach a density so great that a few massive stars would be born. Then, as we see in OB associations, harsh radiation from the massive stars would quickly disperse the cloud, and the speeding stars within would move outward.

Finally, it seems probable that an intermediate range of cloud masses exists for which the two effects are comparable. These clouds would contract at about the same rate at which they lose mass. The result is a molecular cloud containing an ever increasing fraction of young, tightly knit stars but no truly massive ones. Even when stellar winds drive away the cloud, the gravitational attraction among these closely packed stars themselves would be enough to keep them bound for a very long time in a configuration not unlike the one astronomers call an open cluster.

CLOUD CONTRACTION

MY FORCE-BALANCE THEORY described how the starting mass of a parent cloud could determine the interplay of contraction and expansion in, and thus the evolution of, the resulting cluster. Yet although astronomers can observe expansion and dispersal directly in OB associations, no one had found evidence that molecular clouds ever contract at all, let alone in the ways my theory suggested. Such contraction would surely occur in the very first stages of cluster formation, but the youngest stellar groups—the embedded clusters—resisted direct examination. I would have to figure out a way to demonstrate that more mature clusters had undergone contraction long ago.

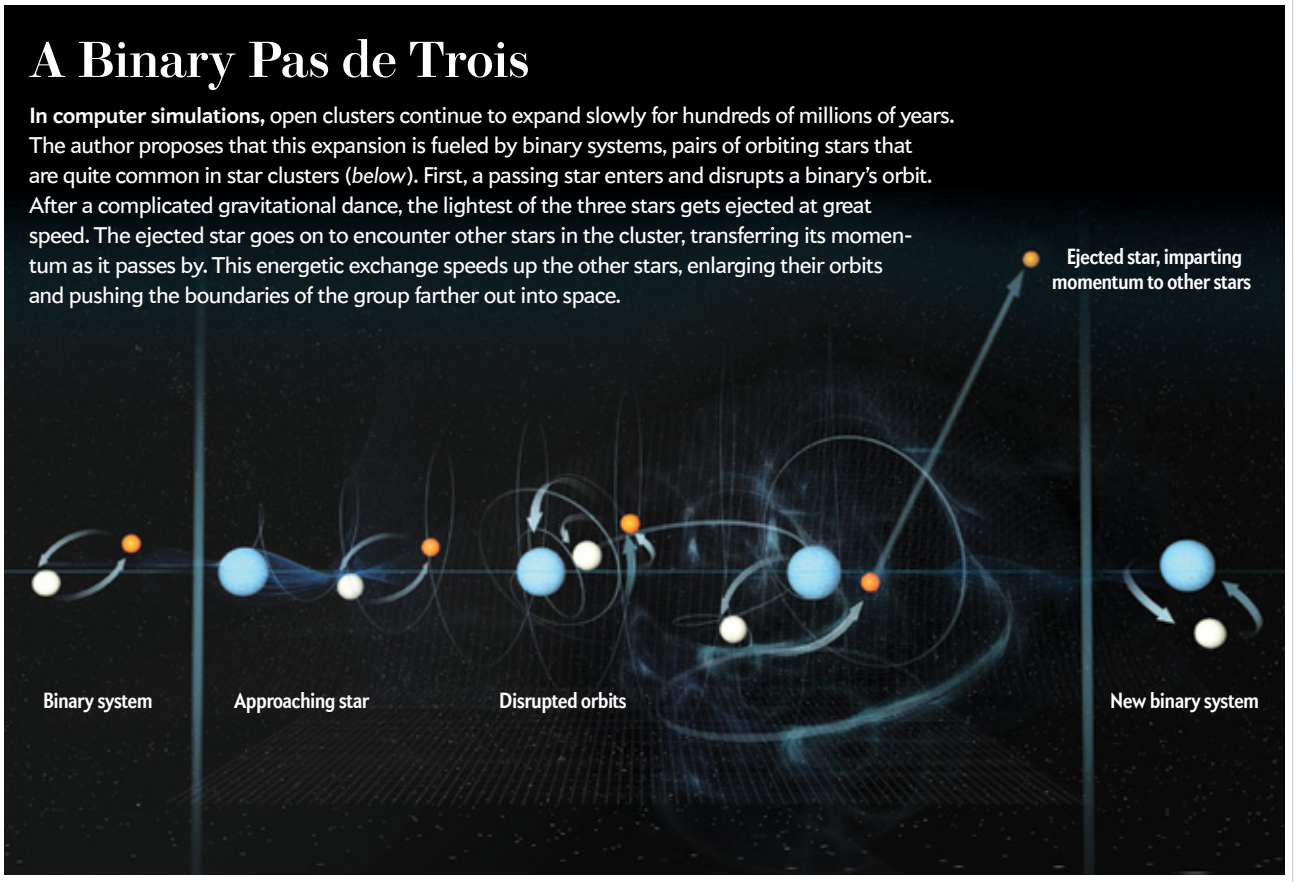
I gained a clue from work done in the late 1950s by astrono-

© 2014 Scientific American

EXPANSION MECHANISM

A Binary Pas de Trois

In computer simulations, open clusters continue to expand slowly for hundreds of millions of years. The author proposes that this expansion is fueled by binary systems, pairs of orbiting stars that are quite common in star clusters (*below*). First, a passing star enters and disrupts a binary's orbit. After a complicated gravitational dance, the lightest of the three stars gets ejected at great speed. The ejected star goes on to encounter other stars in the cluster, transferring its momentum as it passes by. This energetic exchange speeds up the other stars, enlarging their orbits and pushing the boundaries of the group farther out into space.



mer Maarten Schmidt of the California Institute of Technology. Schmidt observed that the birth rate of new stars depends on the density of the surrounding gas. So, I reasoned, if a parent cloud had in fact contracted in the past, its density would have increased, and the rate of star formation would have accelerated, too. My theory therefore posited an acceleration of star formation in the early life of every stellar group.

To test that prediction, I needed to figure out how to measure historical star-formation rates in clusters. Fortunately, the theory of stellar evolution provides a way to do just that. Among many other things, the theory describes how young stars that are not yet burning nuclear fuel—such as T Tauri types—behave over time. T Tauri stars are about as massive as our sun, and they are just as luminous. But instead of shining because of nuclear fusion, they radiate the heat generated by compression as their own gravity causes them to contract. With time, the rate of their compression slows, while their surface temperature climbs. The stars thus get both dimmer and hotter in a predictable pattern as they age.

If you know the surface temperature and luminosity of a T Tauri star, therefore, you can tell how long it has been contracting—in other words, you can determine just how old it is. I realized that the collective set of ages of all these stars in a cluster would reveal the star-formation history of the group—when and at what rate the member stars formed over time.

It was not difficult to apply this method to nearby stellar groups, for which the required data are most readily available. Palla and I found that, for all groups that still possess copious cloud gas, the total star-formation rate has been increasing with time. In 2000, for example, we published data showing that the rate of star formation in the Orion nebula cluster accelerated for millions of years before its parent cloud dispersed. This finding encouraged me to believe that my assumption was correct: early in their history, all cluster-forming clouds probably *do* contract.

In 2007 then graduate student Eric Huff, now at Ohio State University, and I constructed a theoretical model of the parent cloud of the Orion nebula cluster. Our model included the forces of contraction and expansion postulated by my theory. In computer simulations based on the model, the simulated cloud contracted, just as we predicted it would. We then applied an empirical prescription known as the Schmidt-Kennicutt law, derived from Schmidt's observations and many subsequent ones, to infer how the rise of density in a parcel of the cloud over time would affect the local star-formation rate.

Our modeling yielded an accelerating rate of star formation that matched the acceleration Palla and I had derived from ages of stars in the Orion nebula cluster. This additional finding further corroborated the force-balance theory's assumption that parent clouds contract in the early stages of cluster evolution.

CLUSTER EXPANSION

UNFORTUNATELY, the methods I used to measure and model early star-formation rates in clusters such as the Orion nebula

© 2014 Scientific American

group cannot be applied to open clusters, those strangely persistent groups that lack any trace of a parent cloud yet remain bound by gravity. Most open clusters are just too old; their epoch of contraction and star formation—which lasted for just a few million years—amounts to a tiny fraction of these clusters' total lifetimes. The tools for discerning stellar ages do not have nearly the resolution needed. And we cannot yet simulate the parent clouds of open clusters, either; the clouds dispersed so long ago that we cannot even guess at their masses or behavior. So far the early stages of open-cluster evolution remain inaccessible even to indirect observation.

It is possible, however, to model the evolution of an open cluster whose parent cloud has already vanished using so-called N-body simulations. In such simulations, the computer solves the complex, interlocking equations that describe the motion of multiple objects under the influence of their mutual gravitational attraction. This approach has elucidated what happens in open clusters *after* the initial star-forming contraction proposed by my theory and has provided some unexpected insights into the forces that shape cluster expansion.

Although open clusters are remarkably stable, they are not static. The mutual gravity among member stars creates a constant, slow churning, as the orbiting stars weave in and out of one another like bees swarming in a hive. N-body codes describe this gravity-induced dance, and they are so efficient that they can simulate the evolution of a 1,200-member group such as the Pleiades on a standard desktop computer. Several years ago my graduate student Joseph M. Converse, now at the University of California, Berkeley, and I took this numerical path to elucidate the history of the Pleiades. Our strategy was to guess an arbitrary initial configuration for the cluster and then to let it evolve for 125 million years. We compared the simulated cluster with its actual counterpart and changed the initial conditions until the N-body simulation produced a group that resembled the real thing.

What we saw surprised us. It seems that while remaining gravitationally bound, the Pleiades cluster has expanded, more or less uniformly, since its cloud dispersed. The stars in their busy orbits move away from one another at a stately, steady pace. This result conflicts with prior analyses, which had predicted that the stars in open clusters would slowly segregate into an inner clump of heavier ones and an outer envelope of relatively light ones. This pattern of segregation is called dynamical relaxation, and it is the standard description of how gravitationally bound clusters evolve over time. Globular clusters, for example, are known to behave this way. Yet even when we let our N-body simulation run for 900 million years into the future, expansion continued uniformly, showing what an inflated but still intact Pleiades will look like at the age of a billion years.

This finding suggests that the classical analysis overlooked some critical factor in the balance of forces shaping cluster evolution. What drives the uniform expansion of open clusters? Converse and I demonstrated that the key is binary stars: pairs of close, orbiting companions that are quite common in stellar groups. Simulations performed by Douglas Heggie, now at the University of Edinburgh in Scotland, showed in the mid-1970s that when a third star approaches such a pair, the three engage in a complicated dance, after which the lightest of the three is usually ejected at high speed. The ejected star soon encounters other members and shares its energy with them, increasing those stars' orbital velocities and effectively "heating up" the cluster. In our N-body simulations, it was the energy from these binary encounters that caused the open cluster to expand—albeit so slowly that the expansion could easily go unnoticed by astronomers.

ENDURING MYSTERIES

MY INVESTIGATION of star clusters offers some evidence for my proposal that the original mass of a molecular cloud determines both the structure of a cluster and its evolution. The work also offers promising directions for future research. For example, astronomers should look for ways to observe the uniform expansion of open clusters predicted by my studies.

But my findings also serve to highlight the many things we still do not know about star clusters in general. Despite the advances in computer simulations, we do not yet have the tools to model how certain regions of parent clouds become dense enough to form stars. And several decades of radio and infrared observations have failed to reveal the patterns of internal motion in these clouds. The birth phase of stellar groups—a phase that takes place within the thick dust of embedded clusters—remains shrouded in mystery.

Yet the force-balance model my colleagues and I developed can help us figure out more details about this phase and other aspects of cluster evolution. We want to verify, through a combination of analytic studies and N-body simulations, that a cloud losing mass at the same rate as it contracts will indeed produce a gravitationally bound system resembling an open cluster. We also want to use modeling to explore how nascent T associations might reverse the cloud contraction and then disperse into space. Do stellar winds really play the pivotal role astronomers have assumed, for instance?

The impact of this research will extend far beyond the clusters themselves. Although the study of stellar groups in the Milky Way was long a backwater of astronomy, it is fast becoming central to other research. Some astronomers believe, for example, that the sun formed in a crowded OB association and that the close presence of neighboring stars perturbed the surrounding disk of gas and dust in ways that shaped our solar system. The molecular clouds that spawn clusters are also important players in the evolution of the interstellar medium and of galaxies as a whole. Star clusters may thus hold the key to a better understanding of the entire universe: from the birth of our solar system to the past and future of all that exists beyond it. SA

Steven W. Stahler is a theoretical astrophysicist at the University of California, Berkeley, and co-author, with Francesco Palla, of *The Formation of Stars*, the first comprehensive textbook on star creation (Wiley-VCH, 2004).

MORE TO EXPLORE

Embedded Clusters in Molecular Clouds. Charles J. Lada and Elizabeth A. Lada in *Annual Review of Astronomy and Astrophysics*, Vol. 41, pages 57–115; September 2003.

The Orion Nebula: Where Stars Are Born. C. Robert O'Dell. Belknap Press/Harvard University Press, 2003.

The Formation of Stars. Steven W. Stahler and Francesco Palla. Wiley-VCH, 2004.

The Dynamical Evolution of the Pleiades. Joseph M. Converse and Steven W. Stahler in *Monthly Notices of the Royal Astronomical Society*, Vol. 405, pages 666–680; June 2010.

The Birth and Death of Star Clusters in the Milky Way. Steven W. Stahler in *Physics Today*, Vol. 65, No. 10, pages 46–52; October 2012.

scientificamerican.com/magazine/sa

© 2014 Scientific American

STELLAR EVOLUTION

THE FAR, FAR FUTURE OF STARS

Some say its glory days are long gone, but the universe has life in it yet. Brand-new types of celestial phenomena will unfold over the coming billions and trillions of years

By Donald Goldsmith

© 2014 Scientific American

IN BRIEF

Although the grand era of galaxy and star formation is over, the universe remains a vigorous place. In the future stars will gradually shift their appearance as their composition changes. Star and planet systems will fall apart, and celestial objects that now are rare will become common, such as dense balls of helium. In some ways, the universe in the future may be more hospitable to life than it is today. **Considering the far future** of the cosmos is more than inherently interesting. The far distant future provides astrophysicists with an intellectual sandbox, a way for them to grasp the implications of their theories and observations.

Illustration by Kenn Brown, Mondolithic Studios

© 2014 Scientific American

TIME'S SEEMINGLY INEXORABLE MARCH HAS ALWAYS PROVOKED INTEREST in, and speculation about, the far future of the cosmos. The usual picture is grim. Five billion years from now the sun will puff itself into a red giant star and swallow the inner solar system before slowly fading to black. But this temporal frame captures only a tiny portion—in fact, an infinitesimal one—of the entire future. As astronomers look ahead, say, "five hundred and seventy-six thousand million years," as humorist Douglas Adams did in 1980 in *The Restaurant at the End of the Universe,* they meet a cosmos replete with myriad slow fades to oblivion. By then the accelerating expansion of space will have already carried everything outside our galaxy beyond our view, leaving the night sky ever emptier. Lord Byron captured the prospect of such a celestial wasteland in his 1816 poem "Darkness": "The bright sun was extinguish'd, and the stars/Did wander darkling in the eternal space."

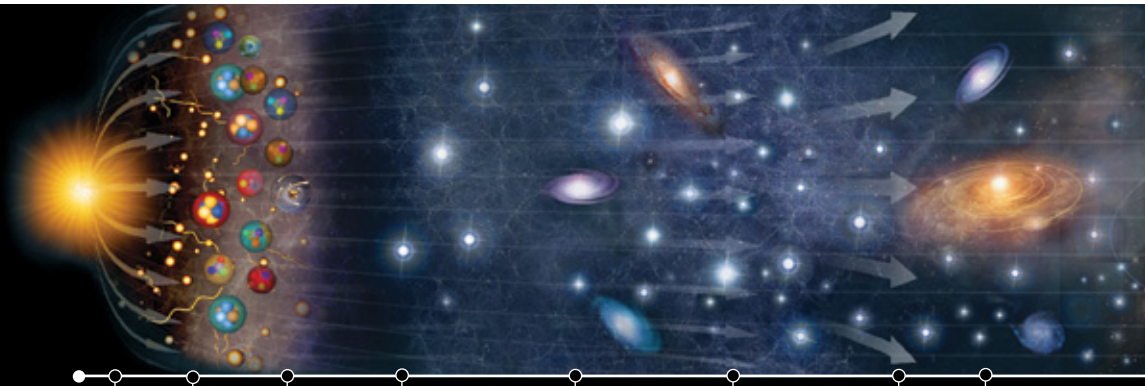
But here's the good news: oncoming darkness captures only half the story. Star formation has indeed long since passed through its most glorious epoch, but the universe has life in it yet. Strange new beasts will enter the astronomers' zoo. Outlandish phenomena that now occur rarely, if at all, will become routine. Cosmic conditions favorable to life may, if anything, become even more abundant.

Scientific eschatology—the study of the far future—has a distinguished history in cosmology and physics. Fascinating in its own right, this endeavor also offers a conceptual testing ground for new theories, plus an opportunity to make abstract ideas more concrete. One of the most abstract ideas of all, the shape of space, may prove easier to grasp when cosmologists describe what it implies for the fate of the universe. Physicists who seek to

COSMIC HISTORY

Dawn to Dusk

After the furies of birth, the mature cosmos now evolves more slowly. Stars will continue to form for as long as another 100 trillion years (about 10,000 times the present age of the universe), which leaves plenty of time for slow-building cosmic phenomena to occur.



Timeline not to scale

Illustrations by Malcolm Godwin

© 2014 Scientific American

reconcile their disparate theories about fundamental particles and forces predict processes that will occur only after trillions of years, or even longer, such as the decay of protons and the evaporation of black holes. Increasingly, astrophysicists bring the far future into their models of stellar and galactic evolution as well. During the past decade they have attempted to reconstruct the ways that the formation and composition of stars and galaxies have changed since the big bang. Their growing knowledge of the past allows them to extrapolate trends into the far future.

FORGOT TO TURN IT OFF

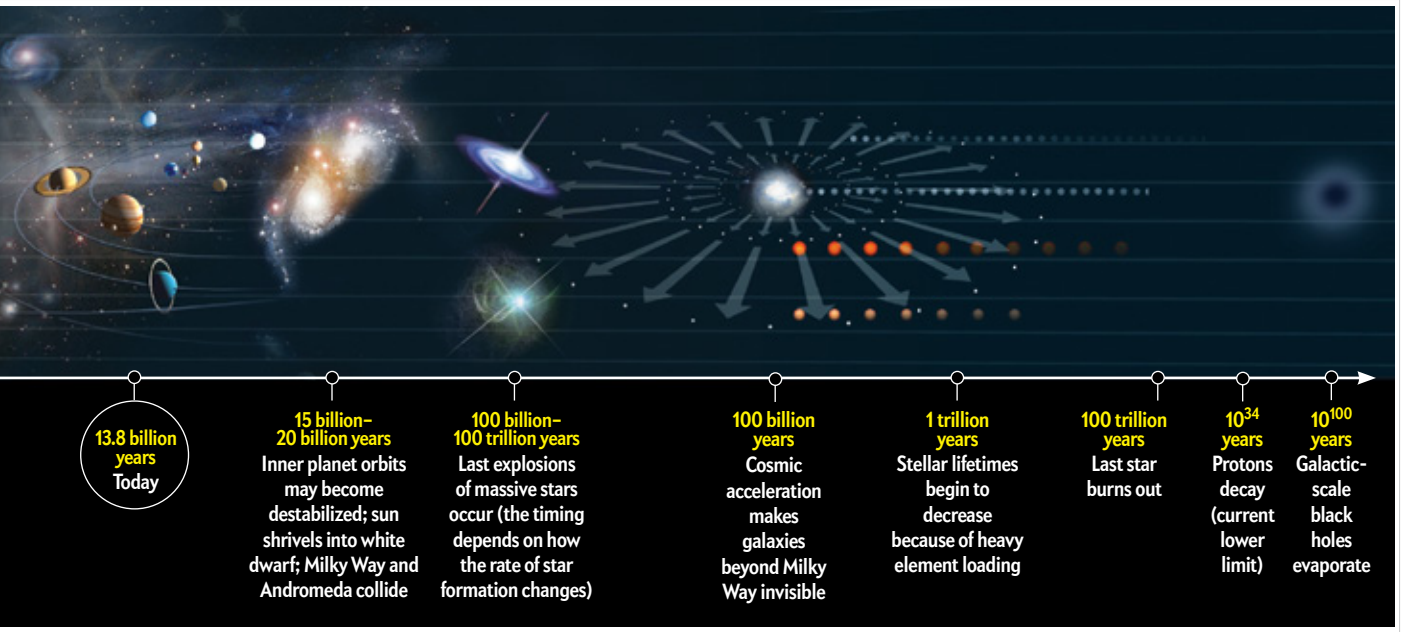
AMONG THIS SUBJECT'S PIONEERS is Greg Laughlin, an expert on star formation at the University of California, Santa Cruz. As a graduate student, he created a computer code to calculate the evolution of extremely low mass stars and forgot to flag it to turn off after reaching the present age of the universe. Left to its own devices, the program ran and ran, producing trillions of years of future predictions—quite wrong, as things turned out, but sufficient to get him hooked on the subject.

To know the future of stars requires understanding how they form. Stars are born within interstellar clouds of gas and dust, which contain hundreds of thousands to several million times the mass of our sun. Such stellar nurseries, sprinkled throughout the Milky Way, gave birth to its few hundred billion stars and will eventually produce tens of billions more. Yet this success consumes the future: the raw material for new stars is being used up. Even though massive stars die in supernova explosions that return some material into interstellar space and even though galaxies can also accrete fresh gas from intergalactic space, the new material cannot replenish all the gas that stars have locked up. The interstellar gas within our galaxy now totals only a tenth or so of the mass in stars.

Today stars form in the Milky Way at a rate close to one solar mass per year, but at its peak, eight billion to 10 billion years ago, the rate was at least 10 times higher. Laughlin estimates that star formation will decrease by approximately a factor of 10 for every factor of 10 in time, so that in 100 billion years it will slow to a tenth of its present rate, and a trillion years from now stars will form at only about one one-hundredth of the current rate.

That said, impressive changes could disrupt the steady march toward stellar obscurity. We in the Milky Way, for example, must soon—"soon," as in a few billion years—confront the arrival of the onrushing Andromeda system, the closest giant spiral galaxy to our own. The dense central regions of these two galaxies will either collide or begin to orbit their common center of mass. Their interaction will produce "Milkomeda." By churning and stirring the interstellar gas and dust, the creation of Milkomeda will reinvigorate star formation temporarily, producing what astronomers call a starburst. Once this growth spurt dies away, the merged system will closely resemble an elliptical galaxy, a mature system with a low density of star-forming material and a consequently low rate of star formation.

In addition to forming in smaller numbers, stars of the future will show the effects of the changes that will occur in their raw material. The fiery furnace of the big bang forged hydrogen, helium and lithium. All the heavier elements have been created by stars themselves, typically late in their lives—either within red giant stars, which shed their outer layers as they age, or during supernova explosions. Red giants provide most of the lighter and more abundant heavy elements, such as carbon, nitrogen and oxygen, whereas supernovae produce a wider range, all the way up to uranium. All of these mix into the existing elemental mulch of interstellar gas, allowing subsequent generations of stars to begin life with more of these materials. The sun, a comparative youngster at five billion years old, has 100 times the heavy element abundance of stars that formed over 10 billion years ago; indeed, some of the oldest stars contain almost no heavy elements at all. Coming generations of stars will be even more enriched, which will alter their inner workings and outward appearance.


13.8 billion years
Today
15 billion–20 billion years
Inner planet orbits may become destabilized; sun shrivels into white dwarf; Milky Way and Andromeda collide
100 billion–100 trillion years
Last explosions of massive stars occur (the timing depends on how the rate of star formation changes)
100 billion years
Cosmic acceleration makes galaxies beyond Milky Way invisible
1 trillion years
Stellar lifetimes begin to decrease because of heavy element loading
100 trillion years
Last star burns out
10^{34} years
Protons decay (current lower limit)
10^{100} years
Galactic-scale black holes evaporate

© 2014 Scientific American

NEW ABODES FOR LIFE

THE STEADILY INCREASING ABUNDANCE of heavy elements within newborn stars produces two noticeable effects. First, it augments the opacity of a star's outer layers. Hydrogen and helium are nearly transparent, but even a modest pinch of heavy elements traps radiation, lowering the star's luminosity. The balance of forces within the star shifts because the lower luminosity means that the star consumes its nuclear fuel at a lower rate. If only this effect were in operation, a star rich in heavy elements would live longer than a star of the same mass that lacks those elements. But a second effect counteracts the first: the heavy elements are nuclear deadweight. Because they do not participate in nuclear fusion, they reduce the amount of nuclear fuel available within a star of a given mass and tend to shorten its life.

Laughlin and his colleague Fred Adams of the University of Michigan made the initial study of these two effects in 1997. They found that the first will dominate for the next trillion years or so, as the increase in heavy elements within new stars raises their opacities and thus lengthens their lifetimes. Eventually, however, the heavy elements will constitute a significant fraction of stars' masses and will begin to shorten their lifetimes. The crossover point will occur when the heavy element fraction within a newborn star reaches about four times the current value.

The extra heavy elements should also favor the birth of planets, along with stars, and thus the prospects for life in the universe. Astronomers have measured the elemental abundances in the stars around which many hundreds of Jupiter-size planets have now been discovered. Their results show that stars with greater heavy element abundances are more likely to have one or more giant planets in orbit around them. "Jovian-type planets show a definite correlation with [heavy element abundances]," says John Johnson, a planet-hunting expert now at Harvard University. "Because the interstellar medium is getting steadily enriched [in heavy elements], planet occurrences will probably increase."

What about Earth-like planets? Recent results from the Kepler space telescope imply that rocky planets are very abundant: nearly one quarter of all sunlike stars appear to have at least one planet similar to Earth in size, orbiting in the star's habitable zone. Yet no correlation seems to exist between the parent stars' heavy element abundances and the likelihood of rocky planets around them. Nevertheless, the universe of the far future should be filled with planets. Despite the diminishing rate of star formation, perhaps half or two thirds of all the planets that will ever exist have yet to be born.

At first, the proliferation of planets does not seem promising for life. Most of the stars of the far future will be much less massive and less luminous than the sun. Fortunately, even a low-mass, dim star can allow life to flourish. A star with as little as one one-thousandth of the sun's luminosity can maintain temperatures that allow liquids to exist on close-in planets, satisfying what seems to be a requirement for living things to exist.

Planets should not only grow generally more common but also be enriched in the stuff of life. In addition to requiring a liquid bath, life on Earth, as well as almost all other forms of life that scientists speculate about, depends on the existence of carbon, nitrogen and oxygen. As time goes on, the increasing relative abundance of these elements should yield planets more hospitable to life. Therefore, as star formation steadily dimin-

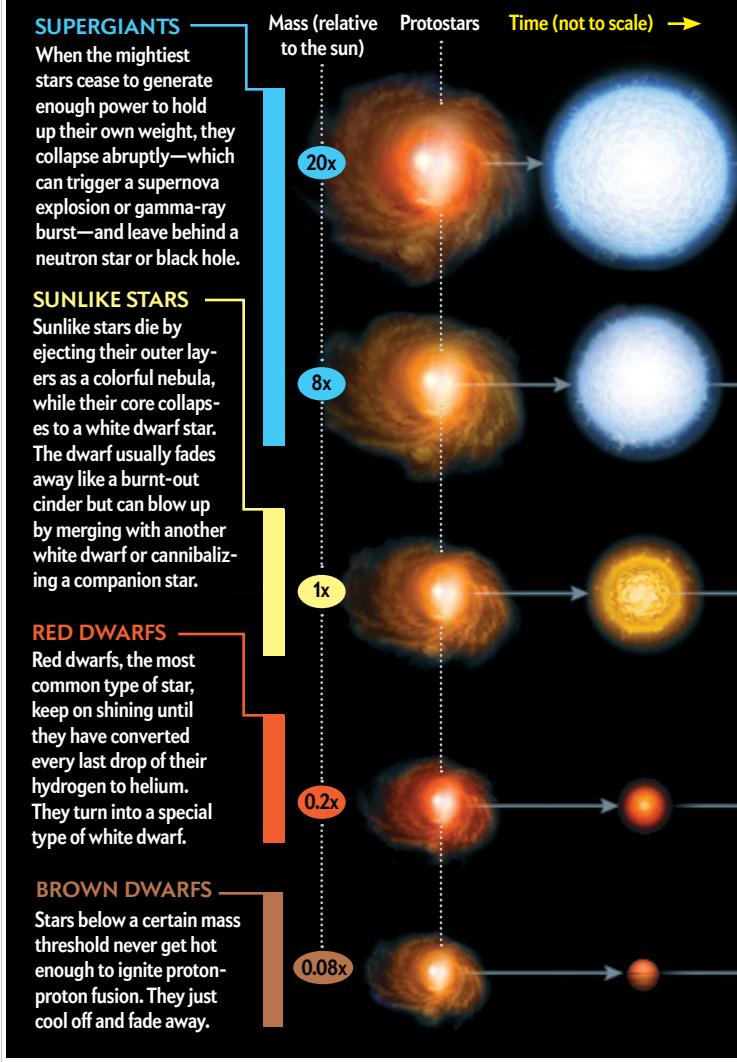
FUTURE OF STELLAR EVOLUTION

The Meek Shall Inherit the Universe

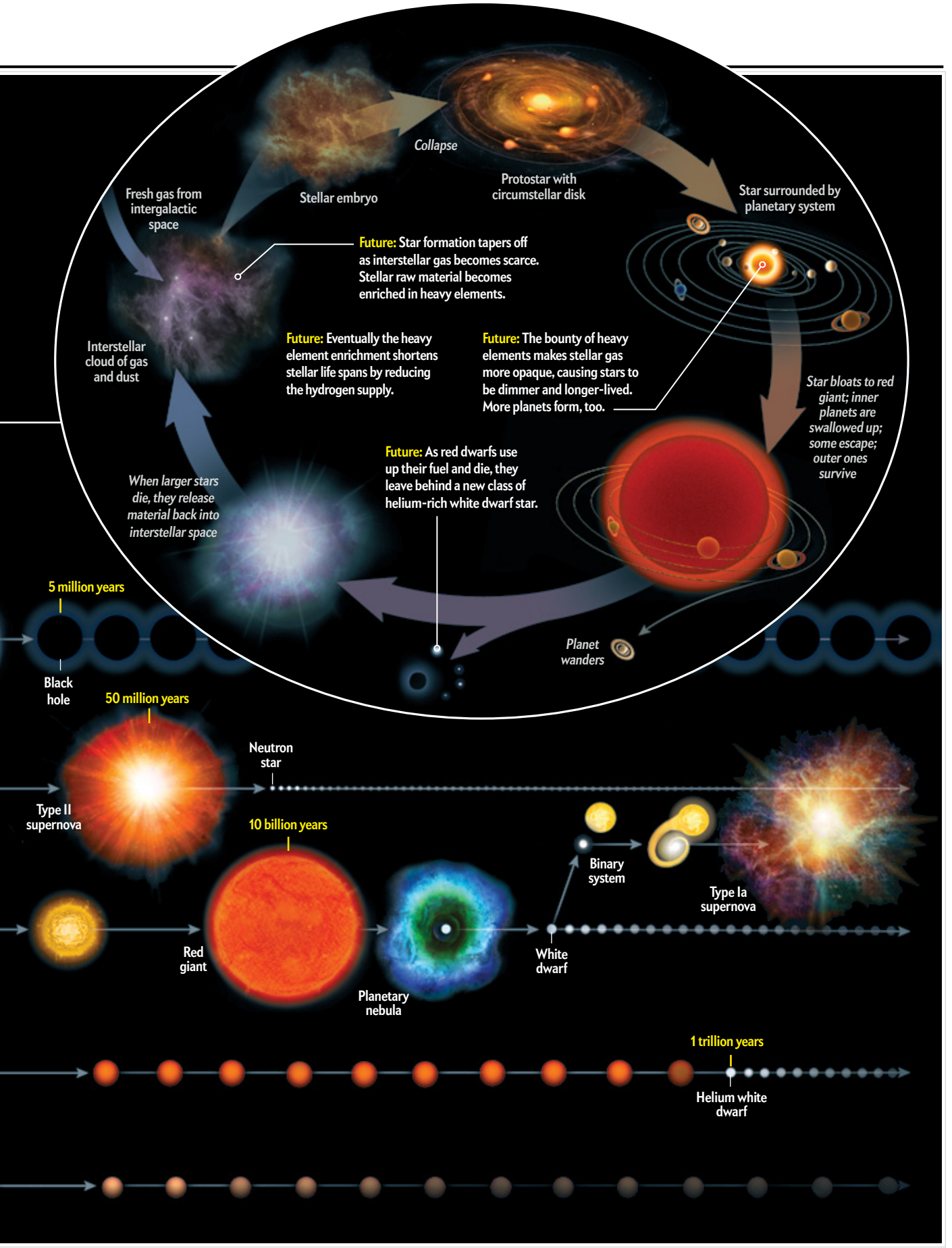
In terms of raw brilliance, the glory days of the cosmos are already behind it. In subtler ways, though, it will remain vibrant for trillions of years to come. Red dwarfs, by far the most common type of star even today, have hardly even begun their life cycles and will eventually develop into novel stellar types. New generations of stars will incorporate the heavy elements forged by their predecessors, changing their appearance and life spans. Planets will, if anything, become even more abundant. Over the vastness of time, rare processes such as direct stellar collisions will become commonplace.

Slow but Steady Wins the Cosmic Race

The life cycle of stars follows a simple rule: the bigger they come, the harder they fall. Massive stars have more fuel but consume it at a disproportionate rate and go out with a bang. Because they live for but a cosmic eyeblink, they rule the galaxy only as long as new ones are continually being born. The future belongs to lesser, longer-lasting stars.


SUPERGIANTS
When the mightiest stars cease to generate enough power to hold up their own weight, they collapse abruptly—which can trigger a supernova explosion or gamma-ray burst—and leave behind a neutron star or black hole.
SUNLIKE STARS
Sunlike stars die by ejecting their outer layers as a colorful nebula, while their core collapses to a white dwarf star. The dwarf usually fades away like a burnt-out cinder but can blow up by merging with another white dwarf or cannibalizing a companion star.
RED DWARFS
Red dwarfs, the most common type of star, keep on shining until they have converted every last drop of their hydrogen to helium. They turn into a special type of white dwarf.
BROWN DWARFS
Stars below a certain mass threshold never get hot enough to ignite proton-proton fusion. They just cool off and fade away.
Mass (relative to the sun)
Protostars
Time (not to scale)
20x
8x
1x
0.2x
0.08x

© 2014 Scientific American


Collapse
Fresh gas from intergalactic space
Stellar embryo
Protostar with circumstellar disk
Star surrounded by planetary system
Future: Star formation tapers off as interstellar gas becomes scarce. Stellar raw material becomes enriched in heavy elements.
Interstellar cloud of gas and dust
Future: Eventually the heavy element enrichment shortens stellar life spans by reducing the hydrogen supply.
Future: The bounty of heavy elements makes stellar gas more opaque, causing stars to be dimmer and longer-lived. More planets form, too.
Star bloats to red giant; inner planets are swallowed up; some escape; outer ones survive
When larger stars die, they release material back into interstellar space
Future: As red dwarfs use up their fuel and die, they leave behind a new class of helium-rich white dwarf star.
Planet wanders
5 million years
Black hole
50 million years
Neutron star
Type II supernova
10 billion years
Binary system
Type Ia supernova
Red giant
Planetary nebula
White dwarf
1 trillion years
Helium white dwarf

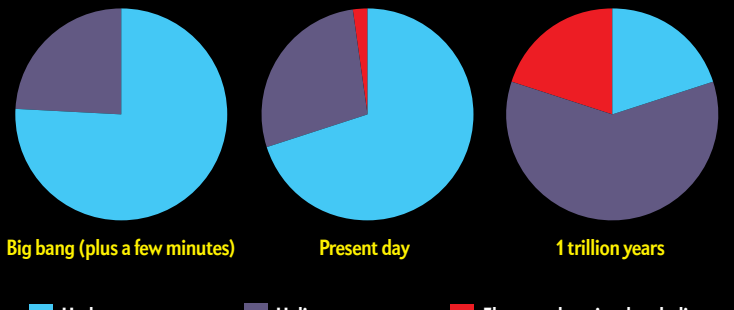
Graphics by Jen Christiansen

© 2014 Scientific American

FUTURE OF STELLAR COMPOSITION

Enriching the Cosmic Mulch

Stars shine by fusing hydrogen into helium and, toward the end of their lives, helium into still heavier elements. Each generation of stars begins life with a greater endowment of heavy elements than the previous one. This process gradually changes the appearance and longevity of stars. It may also increase the number of planets that form.



ishes, every newborn star should appear with a progressively greater probability of lighting one or more potential life-bearing planets. Some of these new stars will have the low masses and tiny luminosities that allow them to last for hundreds or thousands of billions of years (not that such immense lifetimes seem necessary for the origin and evolution of life). However full or empty of life the universe may be today, it should teem with more abundant and more varied forms of life in the future.

WHEN WORLDS COLLIDE

PLANETARY SYSTEMS will endure so far into the future that new considerations will come into play. We take the stability of our solar system for granted; no one worries that Earth's orbit will soon grow chaotic and cause us to collide with Venus. That confidence evaporates when we look to multibillion-year timescales. In 2009 Jacques Laskar and Mickael Gastineau of the Paris Observatory conducted several thousand computer simulations of the future orbits of the sun's four inner planets, varying the planets' initial positions by a tiny amount—just a few meters—between each simulation. They found a probability of about 1 percent that Mercury would smack into Venus during the next five billion years, setting the stage for even more horrific collisions that would probably involve Earth. Over a trillion years such collisions would become highly probable.

The pot will be stirred when the Andromeda galaxy merges with the Milky Way, reconfiguring both galaxies' gravitational fields and possibly triggering the restructuring of the solar system. As Laughlin commented in reviewing Laskar and Gastineau's simulations, "What now remains is to understand the extent to which the hand of dynamical chaos that so lightly touches our solar system has molded the galactic planetary census."

The orbital chaos within a star's planetary family will also occur on much larger scales. The stars in closely bound double-, triple- and higher-multiple star systems orbit the center of mass of each system under their mutual gravitational influences. Much the same is true for star clusters and even entire galaxies. Stars in all these structures almost never make contact; huge expanses of space separate them despite their astronomical neighborliness.

Over long expanses of time, however, "almost never" ratchets up to "sometimes" and ultimately to "almost always." Every double-star system will eventually experience either disruption, as the result of external gravitational forces, or merger, if the two stars orbit so closely that gravitational radiation saps the system of energy. Naturally enough, widely separated double-star systems face the former fate, whereas close-in binaries confront the latter.

When two stars merge, they may temporarily produce a more massive, more luminous star [see "When Stars Collide," by Michael Shara; SCIENTIFIC AMERICAN, November 2002]. Even a planet such as Jupiter can cause a similar effect, though on a smaller scale. Consider a modest star, with just one tenth of the sun's mass and a lifetime close to a trillion years, and suppose it has a Jupiter-like planet. If the planet has an orbital period greater than a few days, it will probably eventually be lost from the system. But if it moves on a tighter orbit, the planet could eventually merge with the star, contributing a fresh supply of hydrogen that would temporarily boost the star's energy output dramatically, producing a novalike outburst. In the future such stellar eruptions will punctuate the slow decline in star numbers and brightnesses. Astronomers even a trillion years from now will observe some strange events among the ever declining numbers of stars in their host galaxies.

LIVE SLOW, DIE OLD

EVEN AFTER TENS AND HUNDREDS of billions of years have elapsed, even when star formation has slowed to a trickle, enormous numbers of stars will continue to shine. Most stars in the universe have low masses and extremely long life expectancies. Stars' lifetimes depend on their masses in a strikingly inverse manner. High-mass stars are so luminous that they burn themselves out quickly and explode after a few million years. Medium-mass stars such as our sun shine modestly and last for billions of years. Stars with significantly less than the sun's mass can endure for hundreds of billions of years or even longer. These stars consume their fuel so slowly that even their meager supplies can feed their nuclear fires through these immense spans of time.

Stars of different masses die in different ways. The sun will become a red giant and, as its outer layers dissipate completely into interstellar space, reveal its core as a white dwarf—a dense, Earth-size stellar corpse made almost entirely of carbon nuclei and electrons. But in stars with less than about 50 percent of the sun's mass, the core temperature never rises high enough to trigger the nuclear reactions that lead to the red giant phase. Instead astronomers think these stars eventually become helium white

© 2014 Scientific American

dwarfs. Such beasts, as the name suggests, consist almost entirely of helium, with little if any hydrogen and just a smattering of other elements. In today's universe they are occasionally born when two close binary stars strip each other's outer layers before they can ignite their helium cores, but astronomers have yet to discover any that arose in the normal course of stellar evolution, because not enough time has passed since the big bang so far. Isolated helium dwarfs are a prime example of a novel phenomenon that our distant descendants (may they live in peace) will one day see for the first time.

Stars with larger masses undergo far more dramatic deaths. The collapse of a massive star core that forms either a neutron star or a black hole triggers a shock wave that blasts the star's overlying layers into space in a supernova explosion. As massive stars disappear from the skies, so, too, will most of these explosions that now punctuate the cosmos. But a second kind of supernova will still occasionally light the skies. This class, called type Ia supernovae, arises in binary-star systems in which one star has become a white dwarf. According to astronomers' most favored models, in some of these stellar pairs hydrogen-rich material from the companion star collects on the white dwarf's surface until its sudden nuclear fusion produces a supernova. Such events will take place as long as there are sufficiently massive companions, perhaps for another 100 billion years or so.

ASTRONOMERS SEE NO SIGN THAT LIFE HAS AFFECTED THE COSMOS ON A GRAND SCALE. THAT WILL CHANGE IN THE FUTURE. ALL THE UNIVERSE WILL BECOME OUR GARDEN.

In a variant of the basic model of type Ia supernovae, two white dwarfs orbit their common center of mass in close proximity. As they do so, their orbital motions cause the binary system to emit gravitational radiation. This emission robs the system of energy and shrinks the size of the white dwarfs' orbits. The approach of the dwarfs proceeds ever more rapidly, until their death spiral melds them in a brief, final paroxysm. Such events might continue to occur for trillions of years.

Even brighter than supernova explosions are gamma-ray bursts (GRBs). These megaexplosions come in two distinct varieties, which apparently originate in two entirely different scenarios. Long GRBs, those whose eruptions of energetic radiation last for two seconds or more, are believed to occur when a massive star's core collapses to form a neutron star. Short GRBs, whose outbursts last for less than two seconds, are thought to result from the merger of a neutron star with either another neutron star or a black hole. Over the coming eons the long variety will become exceedingly rare, as massive stars cease to form, but short bursts might punctuate the heavens for trillions of years.

TRILLIONS AND TRILLIONS

WHEN WE MEASURE COSMIC TIME not by billions but by trillions of years, we enter an epoch when star formation will have ended. All but the lowest-mass stars will have burned themselves out, ending their lives either by blowing up or by withering into white dwarfs. Not counting dark matter, whose composition remains a mystery, our galaxy—and all others in the universe—will then consist primarily of black holes, neutron stars, white dwarfs and extremely faint red stars, so dim that none of them would be visible without a telescope, even at distances less than the current distances from the sun to the nearest stars. How sad, how degenerate, how uninteresting.

And yet, among these dead or fading objects, nature will on occasion produce an enormous outburst, a brief reminder of the nuclear fury that once spangled the heavens with the light from billions of stellar furnaces. If the surviving stars have planets in close propinquity—and we may expect that many or most of them will—then liquid water, along with various forms of life, could appear and endure on their surfaces. Any life that might arise on those planets will have the possibility (already present around the faintest stars) of lasting for epochs well beyond easy imagination, provided they can avoid being blasted into eternity by nearby supernovae or GRBs.

This survey of the far future leaves a great and indeterminate issue. Could highly advanced civilizations, if they exist and persist, change the course of the cosmic history? More than 30 years ago Freeman Dyson of the Institute for Advanced Study in Princeton, N.J., reviewed the situation. The grand leader in this sort of cosmic speculation, he stated, "I think I have shown that there are good scientific reasons for taking seriously the possibility that life and intelligence can succeed in molding this universe of ours to their own purposes." In our present epoch, not even 14 billion years after the big bang, little evidence exists that living things have affected the cosmos on a grand scale. But time's train has barely left the station. In the future the survival of life will require that it commandeer an ever greater fraction of the cosmos's resources [see "The Fate of Life in the Universe," by Lawrence M. Krauss and Glenn D. Starkman; SCIENTIFIC AMERICAN, November 1999]. All of the universe will become our garden.

Bound on this journey for a brief moment, we have little chance of attaining absolute certainty about what will actually happen. Our unfettered minds remain free to roam as far into the future as we choose. As W. H. Auden wrote in his 1957 poem, in an entirely different context: "Were all stars to disappear or die/I should learn to look at an empty sky/And feel its total darkness sublime/Though this might take me a little time." SA

Donald Goldsmith may well be the only astronomer who once worked as a tax attorney. That career was lucrative but short-lived. In 1969 he received his Ph.D. in astronomy from the University of California, Berkeley, and served as a consultant on Carl Sagan's 1980 *Cosmos* series. He was the principal writer for other shows such as NOVA's 1986 *Is Anybody Out There?* with Lily Tomlin and the PBS 1991 series *The Astronomers*.

MORE TO EXPLORE

The Five Ages of the Universe: Inside the Physics of Eternity. Fred Adams and Greg Laughlin. Free Press, 2000.

The Runaway Universe: The Race to Find the Future of the Cosmos. Donald Goldsmith. Basic Books, 2000.

The Formation and Evolution of the Milky Way. Cristina Chiappini in *American Scientist*, Vol. 89, No. 6, pages 506–515; November/December 2001.

An Integrated Picture of Star Formation, Metallicity Evolution, and Galactic Stellar Mass Assembly. L. L. Cowie and A. J. Barger in *Astrophysical Journal*, Vol. 686, No. 1, pages 72–116; October 10, 2008. Preprint available at arxiv.org/abs/0806.3457

scientificamerican.com/magazine/sa

© 2014 Scientific American

EXTRAGALACTIC ASTRONOMY

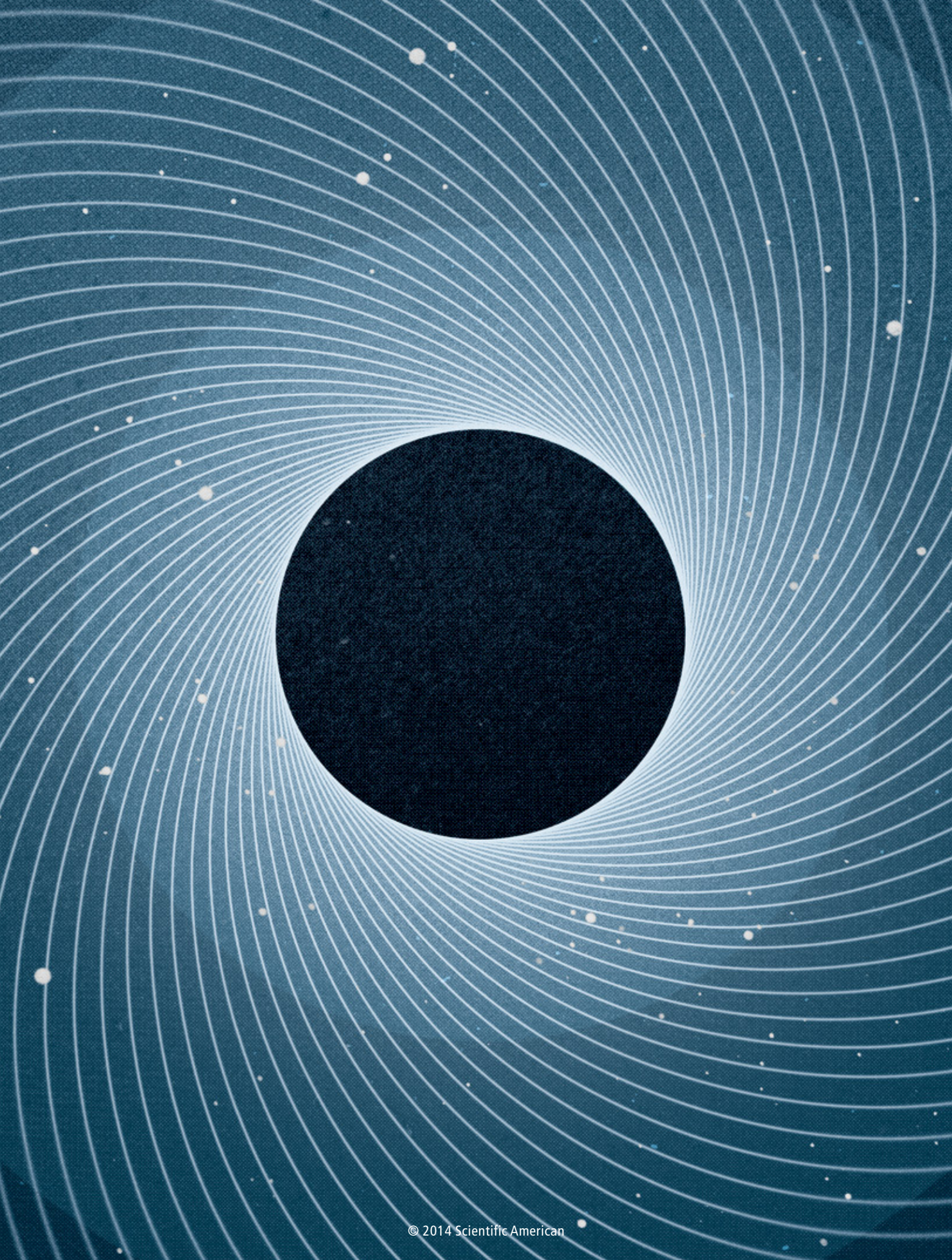
GOLDILOCKS BLACK HOLES

Tipping the scales at less than about a million suns in mass, middleweight black holes may hold clues to how their much larger siblings, and galaxies, first formed

By Jenny E. Greene

Illustrations by Gavin Potenza

© 2014 Scientific American



© 2014 Scientific American

ASTRONOMERS HAVE KNOWN FOR MORE THAN A DECADE THAT NEARLY every large galaxy contains at its core an immense black hole—an object having such intense gravity that even light cannot escape. The death of stars can produce small black holes—with masses ranging from about three to 100 times the mass of the sun—but such stellar-mass black holes are tiny compared with the behemoths at the centers of galaxies, measuring millions to billions of solar masses.

These supermassive black holes pose major puzzles: Why are they so common in galaxies? Which came first—the galaxy or the hole? And how did they form in the first place?

The mystery is intensified because supermassive black holes were already in place when the universe was very young. In June 2011, for instance, astronomers reported the earliest example detected so far—a hole of about two billion solar masses that existed 13 billion years ago, only about 800 million years after the big bang. How could black holes get so big so quickly?

Such rapid formation is perplexing because although black holes have a reputation as mighty vacuum cleaners, they can also act like immense leaf blowers. Gas falling toward a black hole ends up swirling around the hole in a huge disk, the so-called accretion disk. The material heats up and emits radiation, particularly as it approaches the point of no return at the inner margins of the disk. The radiation pushes away other infalling material, limiting how fast the hole can ordinarily grow by accretion. Physicists calculate that a black hole sucking in surrounding matter continuously at its maximal rate would double its mass every 50 million years. That is too slow for a "seed" black hole of stellar mass to grow into a billion-sun monster in less than a billion years.

Astrophysicists have proposed two general ways for seed black holes to form. The first, considered for many years, assumes that the earliest huge black holes were indeed the remnants of stars. The first stars ever to have taken shape in the universe were likely to have been extremely massive compared with those that came later, such as our sun, because the primordial gas clouds were free of elements that help the gas to cool and form smaller clumps. These big stars would have burned out fast and produced black holes of perhaps 100 times the mass of the sun. Some process must then bulk up these holes faster than ordinary accretion. For example, if a big hole formed in a dense cluster of stars, it would end up near the cluster center, with other massive stars and black holes. It could then quickly grow to 10,000 solar masses by swallowing other holes and thereby beating the usual feeding limits. Further growth to supermassive scale could be by more ordinary accretion, perhaps with other largish holes on the menu as well.

Once astronomers knew that large supermassive black holes existed very early on, though, they began to wonder if stellar-mass holes could become supermassive quickly enough, even beginning life with this kind of accelerated growth. People started to look for alternative ways to produce seed black holes, routes that would generate bigger holes than those that could form in the death throes of stars.

Researchers proposed models that made bigger seeds by skipping the middlemen (stars). Rather a large cloud of gas would collapse to create a black hole directly, one larger than the product of a dead star. By making seeds with masses of 10,000 to

IN BRIEF

Black holes with a billion times the sun's mass already existed early in the universe. How did these behemoths grow so big, so fast? What process formed the "seed" black holes from which they grew?

Did the death throes of the first stars provide numerous small seeds that then merged together, or did vast primordial gas clouds bypass the star stage and collapse to form larger seeds directly?

Astronomers are trying to solve this mystery by finding and analyzing leftover seeds—"middleweight" black holes. Early indications suggest that middleweights formed by direct collapse.

© 2014 Scientific American

100,000 suns, this process somewhat alleviates the time crunch to form supermassive black holes at early times. This kind of direct collapse does not happen in our universe today, but conditions were different when the universe was young.

Unfortunately, it is hard to figure out which of these two scenarios took place—whether seed black holes started off small, as the products of dying stars, or on the contrary, started off larger, as the products of gas implosions. Although astronomers can peer far back in time by looking out to vast distances with telescopes, they cannot yet hope to detect seed holes in the act of forming; even the biggest seeds would be too small to be seen so far away. (The James Webb Space Telescope could reveal them, but it is not due for launch until 2018.) So my colleagues and I have been pursuing another strategy: looking for leftover seeds that have, for whatever reason, survived to the present day without growing supermassive.

If seed black holes started off as stars, we would expect to find many leftover seeds, in both the centers and outskirts of galaxies, because primordial stars could have died anywhere in the galaxy. We would also expect to find a continuous range of masses from 100 to 100,000 suns because their growth could get interrupted for lack of food at any stage along the way to supermassive status. In contrast, if seeds formed mainly by direct gas collapse, then the leftovers should be pretty rare; the direct collapse process, if it happened at all, would have occurred less often than ordinary star death. Instead of a wide range of masses, we would find most leftover seed black holes to be heavier than 100,000 suns (the theoretical models indicate that is likely the typical mass of the seeds that would form by direct gas collapse).

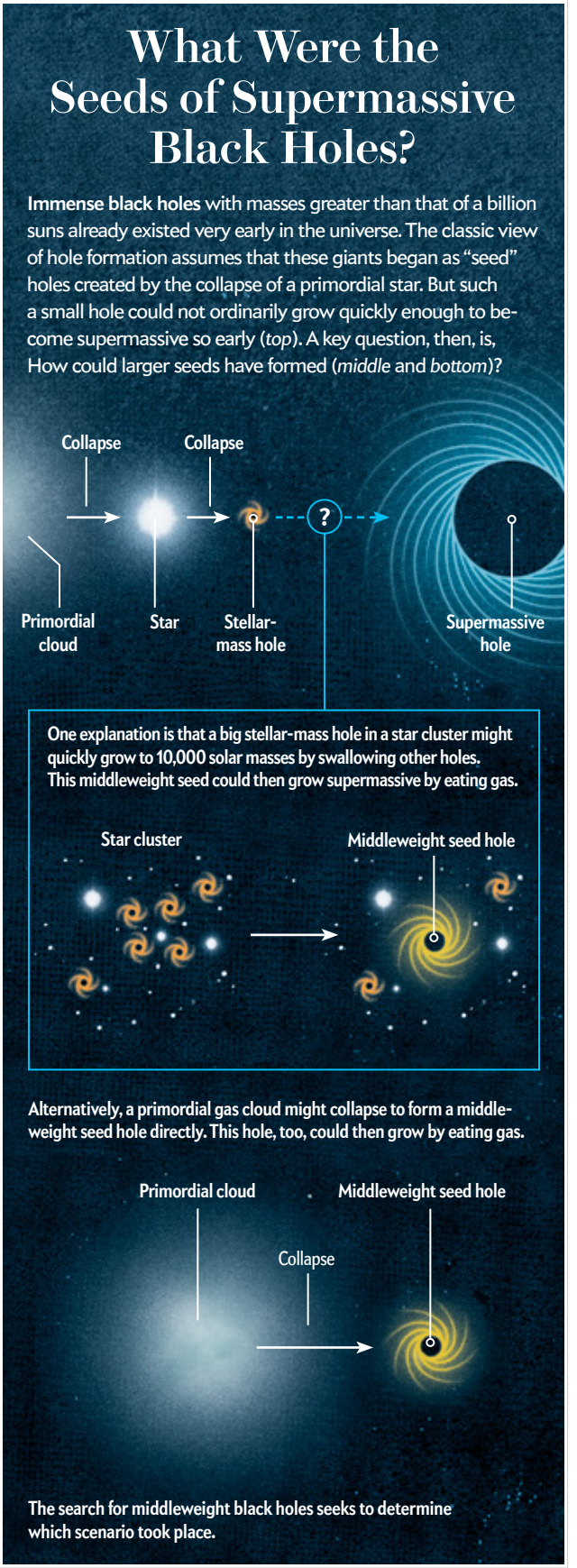
Other astronomers and I have therefore been scouring the skies for a new type of black hole, neither stellar in mass nor supermassive but somewhere in between: the so-called intermediate-mass, or middleweight, black hole. Our aim is to see if the prevalence and range of sizes are more consistent with the star collapse or gas collapse models. When we began this effort about a decade ago, it did not look promising. Astronomers knew of only one middleweight hole and considered it to be a fluke. Since then, though, we have found hundreds.

What counts as a "middleweight"? Here I will take it to mean a black hole with an estimated mass between 1,000 and two million suns. That upper limit is somewhat arbitrary, but it excludes the smallest well-known supermassives, such as the Milky Way's four-million-sun hole. In any case, the boundary is inherently fuzzy. In practice, measurements of a hole's mass often start out very uncertain—for instance, the masses of our first batch of middleweights all shifted up by about a factor of 2 a few years ago when we improved our measurement technique. The precise boundary does not matter as long as we study the entire population of holes extending down from the low supermassive range. What we have learned so far has already provided a new view of the interactions between black holes and the galaxies that they live in.

ELUSIVE MIDDLEWEIGHTS

BLACK HOLES can reveal themselves in a number of ways. For instance, stars whipping around orbits at the very center of a galaxy are a telltale sign of a lurking supermassive black hole. Middleweight holes, however, are too puny to give away their presence by their gravity in this way. Instead we focus on "active"

THE UNDERLYING QUESTION


What Were the Seeds of Supermassive Black Holes?
Immense black holes with masses greater than that of a billion suns already existed very early in the universe. The classic view of hole formation assumes that these giants began as "seed" holes created by the collapse of a primordial star. But such a small hole could not ordinarily grow quickly enough to become supermassive so early (top). A key question, then, is, How could larger seeds have formed (middle and bottom)?
Collapse
Collapse
?
Primordial cloud
Star
Stellar-mass hole
Supermassive hole
One explanation is that a big stellar-mass hole in a star cluster might quickly grow to 10,000 solar masses by swallowing other holes. This middleweight seed could then grow supermassive by eating gas.
Star cluster
Middleweight seed hole
Alternatively, a primordial gas cloud might collapse to form a middleweight seed hole directly. This hole, too, could then grow by eating gas.
Primordial cloud
Middleweight seed hole
Collapse
The search for middleweight black holes seeks to determine which scenario took place.

© 2014 Scientific American

A FIELD GUIDE

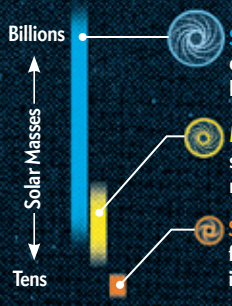
Where Different Kinds of Black Holes Live

Galaxies come in many types, some of which routinely contain supermassive black holes. Our galaxy, the Milky Way (*left*), is a disk, or spiral, galaxy with a bulge (a large, dense collection of stars) and has a supermassive black hole of four million solar masses at its core (*blue*). Many stellar-mass black holes (*orange*) have also been detected in the Milky Way.

Bulged galaxies and large elliptical galaxies (*center*) all seem to harbor a supermassive hole at their core. In contrast, middleweight black holes (*right, yellow*) are more common in galaxies without a large bulge, such as bulgeless disks. Stellar-mass black holes occur throughout galaxies of all types.

Classes of Black Hole

Known black holes fall into three classes, based on their mass, which is usually given in terms of solar masses, or "suns."

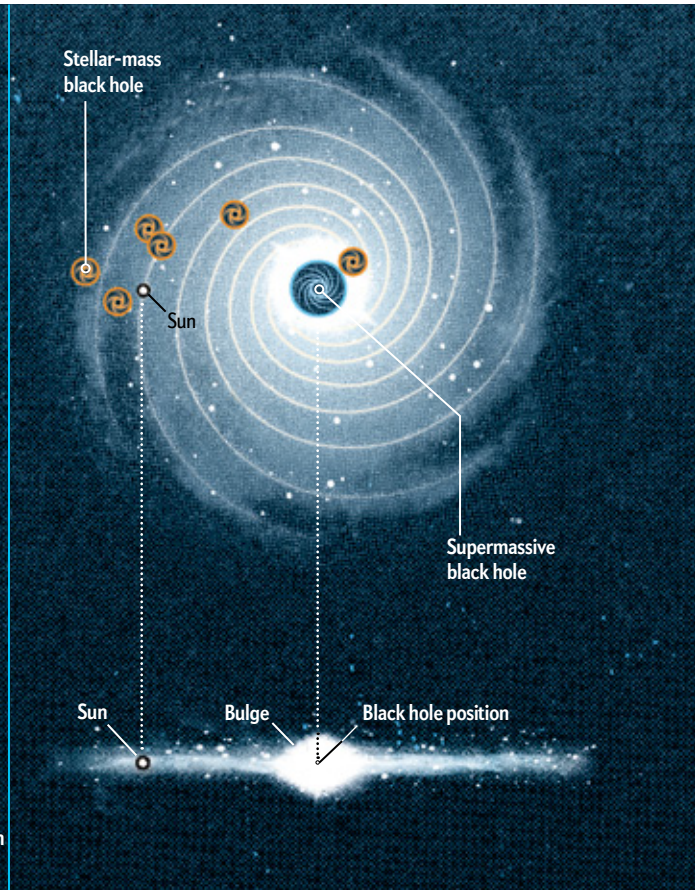


SUPERMASSIVE BLACK HOLES: Millions to billions of suns. Found in the core of all big elliptical or bulged galaxies.

MIDDLEWEIGHT BLACK HOLES: 1,000 to two million suns. Suspected of being "leftover" seed holes that never finished growing to full supermassive size.

STELLAR-MASS BLACK HOLES: Known examples range from four to 30 suns. Formed when large stars collapse, which in theory can produce a hole of three to 100 solar masses.

Black holes are not drawn to scale



Disk galaxy with bulge (such as the Milky Way) and supermassive black hole

THE UNIVERSE

black holes—ones that happen to be eating stuff—because the hot infalling material emits a tremendous amount of light.

Over decades of studies astronomers found that active black holes usually live in big galaxies of a certain kind. Galaxies, particularly massive ones, come in two general types. Some, like our own, have a large, rotating disk of stars. These disk galaxies look like dinner plates when seen edge-on. The other kind, elliptical galaxies, are basically balls of stars. Some disk galaxies actually have small elliptical galaxies at their centers, known as bulges. Active black holes are most commonly found in large elliptical galaxies and disk galaxies with healthy bulges. Nearly every bulge astronomers look at that is close enough to tell turns out to harbor a black hole of millions to billions of solar masses. Furthermore, bigger bulges have bigger holes—the black hole's mass is usually about one 1,000th of the bulge's mass. This surprising correlation is a mystery in its own right, implying that galaxies and supermassive black holes evolve together in ways that astrophysicists have not yet fathomed. More prosaically, this pattern suggested where to look for middleweight holes: in the smallest galaxies. But which ones?

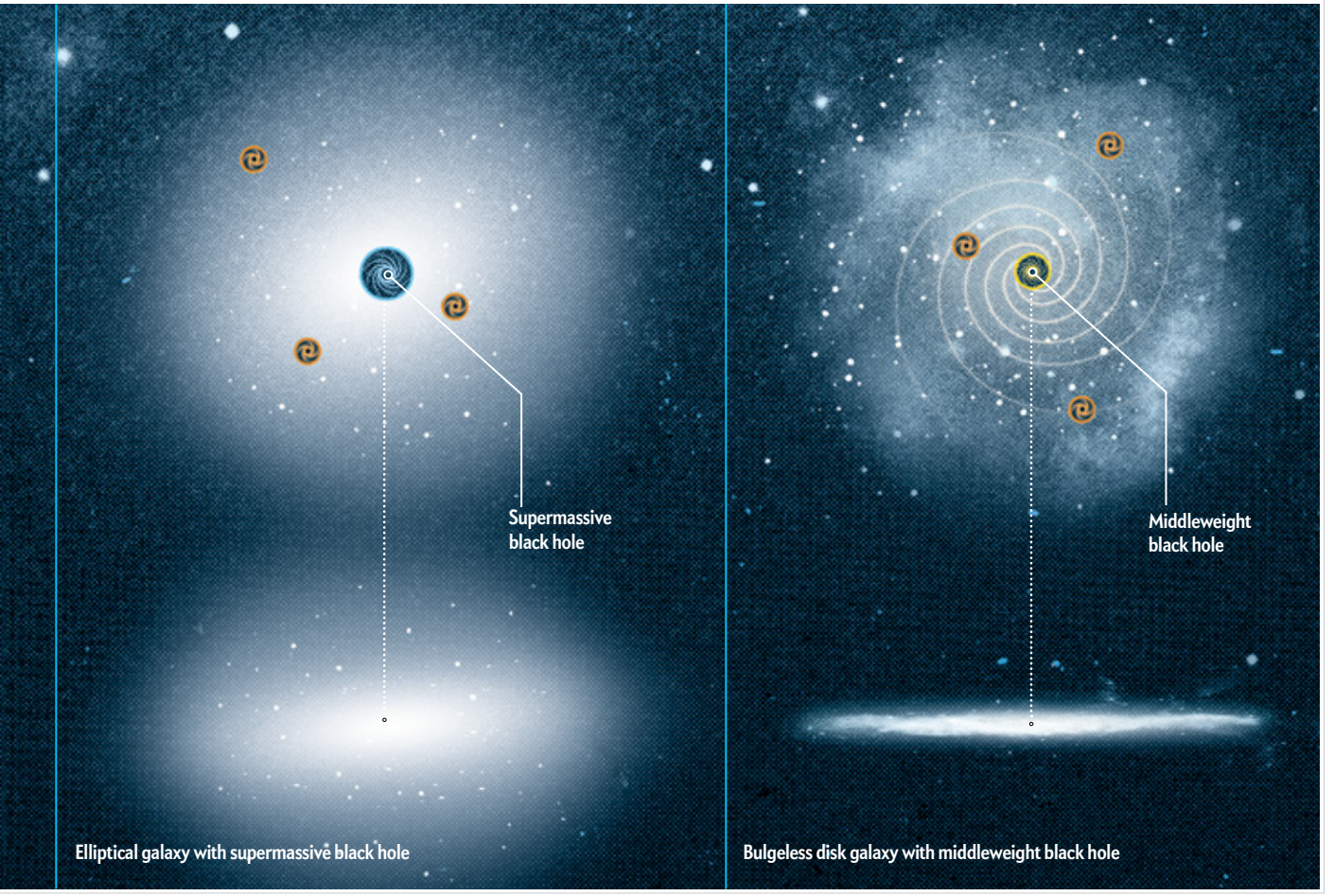
A very puzzling little galaxy offered one idea. My thesis adviser, Luis C. Ho of the Carnegie Observatories, studied about 500 of the nearest bright galaxies for his own thesis in 1995. He found that while most of the galaxies with big bulges contain active black holes, galaxies without bulges do not—with one interesting exception. NGC 4395 is a disk galaxy with an active black hole and no bulge at all. Ho's own thesis advisers had noted this oddity as long ago as 1989, but most researchers considered it an anomaly. Except for NGC 4395, Ho's survey confirmed the broader rule: black holes are not found in bulgeless galaxies.

Accurately estimating the mass of the hole in NGC 4395 is a challenge. The most direct mass measurements in astronomy involve measuring orbital motion. For instance, the speed of a planet and the size of its orbit around the sun let us calculate the sun's mass. Similarly, the orbits of stars in a galaxy can reveal a black hole's mass but only if it is large enough for the effects of its gravity to be discernible in astronomers' observations of the star motions. The hole in NGC 4395 is too small.

Astronomers must therefore rely on less direct clues. For instance, x-rays coming from active black holes change in intensity over time, and the larger the black hole, the more slowly these variations occur. In 2003 David C. Shih and his colleagues, then at the University of Cambridge, found that the intensity of x-rays coming from NGC 4395 varies so quickly that it must be relatively small—most likely 10,000 to 100,000 solar masses. Ho arrived at the same rough mass range based on other evidence, also in 2003.

A slightly more direct measurement of the mass came in 2005, from Bradley M. Peterson of Ohio State University and his co-workers. They used the Hubble Space Telescope and a tech-

© 2014 Scientific American


Supermassive black hole
Middleweight black hole
Elliptical galaxy with supermassive black hole
Bulgeless disk galaxy with middleweight black hole

nique called reverberation mapping, which relies on gas clouds orbiting the hole similar to using planets orbiting the sun. The timing of echoes of light from the clouds provides the size of the orbits. Peterson and company concluded that the black hole is about 360,000 solar masses. Even with this technique, however, the mass has a large uncertainty—as much as a factor of 3—because of assumptions that feed into the number crunching.

The bulgeless galaxy NGC 4395 appears to host just the kind of intermediate-mass black hole we were looking for. Yet of the 500 galaxies examined by Ho, it was the only bulgeless one with clear evidence for an active black hole. The second was found in 2002. Aaron J. Barth, then at the California Institute of Technology, used the Keck II telescope in Hawaii to take a spectrum of a peculiar but little-studied galaxy called POX 52. Like NGC 4395, this galaxy had shown some signs of an active hole even though it is not one of the usual suspects for harboring a supermassive black hole (it is a rare type known as a spheroidal, which is distinct from the bulged disk and elliptical galaxies).

Barth sent the new POX 52 spectrum to Ho, who took one look at it and immediately asked Barth, "Where did you find such a beautiful spectrum of NGC 4395?" The two objects' spectra looked so similar that Ho could not tell them apart. (Features in the spectrum are what indicate the presence of a black hole.)

Because POX 52 is 300 million light-years distant (20 times farther away than NGC 4395), astronomers' mass estimates for its black hole are considerably less direct. Still, a variety of evidence all indicates that the galaxy harbors a black hole of around 100,000 suns. Middleweight black holes in bulgeless galaxies now formed a class of two.

Of course, to solve the bigger problem of how the seeds of supermassive black holes formed, we needed more middleweight specimens to answer a lot of basic questions: How common are middleweight black holes? Does every bulgeless galaxy contain one, or are most such galaxies holeless? Do these middleweight holes occur anywhere else? And are there specimens even smaller than these first two waiting to be found? Only by answering these questions could we learn about how seed black holes formed and what role they played in the early universe.

COMBING FOR HOLES

UNFORTUNATELY, astronomers' standard techniques are biased against finding active middleweight black holes. The larger the black hole, the more it can eat and the brighter it can shine. Smallish black holes are faint and therefore harder to find. But it gets worse. The elliptical galaxies where large black holes tend to occur are extremely well behaved. These galaxies do not have much gas and are not making new stars, leaving a clean and unobstructed view of the galaxy center. In contrast, disk-dominated galaxies (where we suspected middleweight black holes might commonly lurk) are often forming stars, and

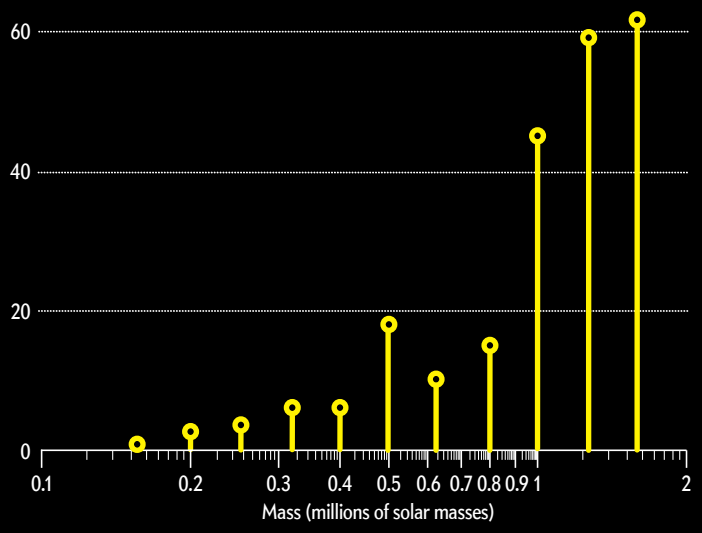
© 2014 Scientific American

MIDDLEWEIGHTS FOUND SO FAR

Early Evidence Favors Gas Collapse over Star Mergers

Analysis of optical light recorded from 500,000 galaxies has turned up more than 100 black holes with masses estimated to be under two million suns (*graph*). Other searches, using midinfrared light, x-rays and radio wavelengths, have detected additional candidate holes. So far it seems that most bulgeless galaxies do not have a middleweight at their core. These observations lend support to the direct collapse scenario of seed formation. If star collapse accounted for the earliest seeds, scientists would expect to have found many more smaller holes in the 10,000- to million-sun range.

Number of Middleweight Black Holes Detected by Optical Light Analysis



THE UNIVERSE

the young starlight and associated gas and dust can hide the active hole.

To overcome these obstacles, in 2004 Ho and I turned to an invaluable library of data designed for finding needles in the cosmic haystack—the Sloan Digital Sky Survey. Since 2000 this project's dedicated telescope in New Mexico has snapped images across more than a quarter of the sky and has recorded the spectra of millions of individual stars and galaxies.

We combed through 200,000 galaxy spectra and found 19 new candidates similar to NGC 4395—small galaxies that contained active black holes with masses we estimated at less than a million suns. Similar searches over the past few years, using more recent Sloan survey data, have expanded the total to about three dozen holes with masses under a million suns and more than 100 just over the million-sun threshold.

The method used to estimate these masses is relatively indirect. The Sloan light spectra tell us the speed of hot gas orbiting a hole. That is only half the information needed to compute the hole's mass directly (the other half is orbit-size). Still, astronomers know from observing active holes in the million- to billion-sun range how the gas speed usually trends with the hole's mass (smaller hole means slower gas). Extrapolating to somewhat lower-mass holes lets us pick out our little guys from the Sloan data.

These searches confirmed what we expected based on NGC 4395 and POX 52: a wider population of intermediate-mass black holes exists. Also in line with expectations, they are found preferentially in galaxies without bulges. Yet these holes still seem to be very rare. Only one in every 2,000 of the galaxies bright enough to study in the Sloan survey shows evidence for an active intermediate-mass black hole.

The Sloan searches, however, could be missing many black holes. They rely exclusively on optical light (the range of wavelengths our eyes can see), and dust clouds could well be hiding many black holes from sight. To get around this, astronomers are using wavelengths of light that can pass right through dust, such as x-ray, radio and midinfrared. Shobita Satyapal of George Mason University and her collaborators have been using midinfrared light to look for signs of hidden active black holes in bulgeless galaxies. Extreme ultraviolet light coming from material plunging into an active hole would wreak havoc in the surrounding gas, creating unusual species, such as excited states of highly ionized neon. Emissions from these ions would leave characteristic fingerprints in the midinfrared spectra. Relatively few galaxies are amenable to this kind of search, and Satyapal's team has found only a couple of new active middleweight black holes. Astronomers have also seen signs of possible middleweight or smallish supermassive black holes at x-ray and radio wavelengths, and follow-up observations to confirm these candidates continue.

These results indicate that the optical searches are indeed overlooking numerous bulgeless galaxies that hide their middleweight holes behind dust—but not enough to make middleweight holes common. The verdict is still out, but perhaps only 5 to 25 percent of bulgeless galaxies harbor a middleweight hole big enough to detect.

GROWING GALAXIES AND HOLES

THE OBSERVATIONS of middleweight holes in bulgeless galaxies may help explain the connection between larger holes and big bulges. As I mentioned earlier, supermassive holes in massive bulged galaxies tend to be about one 1,000th of the mass of the bulge. The growth of a supermassive black hole appears to be intimately linked with the growth of the surrounding bulge. If the correlations between black holes and galaxies are established during the formation of the bulge, then there should be no correlations between the properties of bulgeless galaxies and their middleweight black holes.

A leading theory to explain how this tight correlation comes about in bulged galaxies goes like this: Elliptical galaxies and large bulges form when disk galaxies merge. During the merg-

SOURCE: "A NEW SAMPLE OF LOW-MASS BLACK HOLES IN ACTIVE GALAXIES," BY JENNY E. GREENE AND LUIS C. HO, IN *ASTROPHYSICAL JOURNAL*, VOL. 670, NO. 1; NOVEMBER 20, 2007

Graphic by Jen Christiansen

© 2014 Scientific American

GALAXY NGC 4395, a bulgeless disk galaxy, was the first to show signs of having a middleweight black hole at its core.

COURTESY OF BOB FRANKE *Focal Ponte Observatory*

er, gravitational forces stir up the disks, so the stars no longer orbit in a disk but move around randomly in a ball (the new elliptical or bulge shape). Gas clouds collide during the merger and are funneled toward the center of the bulge, triggering a major burst of star formation, which increases the total mass of stars in the bulge. At the same time, the black holes from each galaxy merge together and eat some of the new gas in the galaxy center. In this way, large bulges and supermassive black holes can grow and evolve together through these large-scale processes that occur in galaxy mergers. By the time the hole reaches about one 1,000th of the bulge mass, its leaf-blower aspect comes to the fore, pushing the remaining gas out of the galaxy center and ending the growth spurt.

Middleweights in bulgeless galaxies, such as NGC 4395, would never have had the benefit of these organized feasts. Instead they would be leftover seeds that have grown only by more happenstance meals of gas at the galaxy center—snacks that are not connected with events shaping the overall evolution of the galaxy. Some galaxies without bulges may not grow a black hole at all. That is the case for the pure disk galaxy M33 (a galaxy much like NGC 4395 in physical appearance), which very clearly contains no black hole more massive than 1,500 suns. Evidence is mounting for this picture, linking black hole growth with bulge formation, but many details remain to be worked out, and the case is not completely settled.

On the question of how black hole seeds formed in the first place, the rarity of middleweight holes lends weight to the theory of direct collapse of gas clouds in the early universe. If star collapse accounted for the earliest seeds, we would expect almost all those galaxies to contain a black hole of at least 10,000 suns at their center. It seems, however, that most small bulgeless galaxies do not contain such a hole at their center.

Other evidence also points toward the direct collapse scenario. In particular, the weak correlation of the middleweights' masses with their host galaxy masses more closely resembles that scenario's predictions. And it is much easier to make a billion-sun hole in a few hundred million years if the seeds start out heavy.

Of course, as more data come in, the conclusions drawn so far could change. For instance, if astronomers were to look at galaxies slightly fainter than those with spectra in the Sloan survey, the fraction of galaxies with middleweight holes might rise or fall. And it is possible that some galaxies contain middleweight black holes outside of galactic centers. Indeed, the search for middleweight holes is continuing on many fronts, as is described in detail at blogs.scientificamerican.com/life-unbounded/2012/09/20/black-hole-roundup.

For now many critical questions about middleweight black holes remain open. Are they more common in specific types of small galaxies? (Such correlations might suggest new ways that holes and their host galaxies interact even before the merges that generate bulges and supermassive holes.) Do most bulgeless galaxies completely lack a middleweight hole, or do they have black holes just slightly too small to be detected so far—perhaps in the range of 1,000 solar masses? (Such holes would surely have grown from remnants of dead stars and *not* formed by direct gas collapse.) Or do all bulgeless galaxies have hefty 10,000- to 100,000-sun black holes, although most of them do not happen to be eating and spewing out x-rays and light? (That would change the conclusion that middleweights are rare.) The answers could push astrophysicists' theories of how galaxies and black hole seeds first formed in radically different directions. SA

Jenny E. Greene pioneered the study of lower-mass black holes in galaxy centers as a part of her Ph.D. thesis at Harvard University. Now an assistant professor in astronomy at Princeton University, she investigates the evolution of galaxy structure more generally. She also teaches algebra to inmates in New Jersey prisons.

MORE TO EXPLORE

A Low-Mass Central Black Hole in the Bulgeless Seyfert 1 Galaxy NGC 4395. Alexei V. Filippenko and Luis C. Ho in *Astrophysical Journal,* Vol. 588, No. 1, pages L13–L16; May 1, 2003.

Active Galactic Nuclei with Candidate Intermediate-Mass Black Holes. Jenny E. Greene and Luis C. Ho in *Astrophysical Journal,* Vol. 610, No. 2, pages 722–736; August 1, 2004.

Nuclear Activity in Nearby Galaxies. Luis C. Ho in *Annual Review of Astronomy and Astrophysics,* Vol. 46, pages 475–539; September 2008. http://arxiv.org/abs/0803.2268

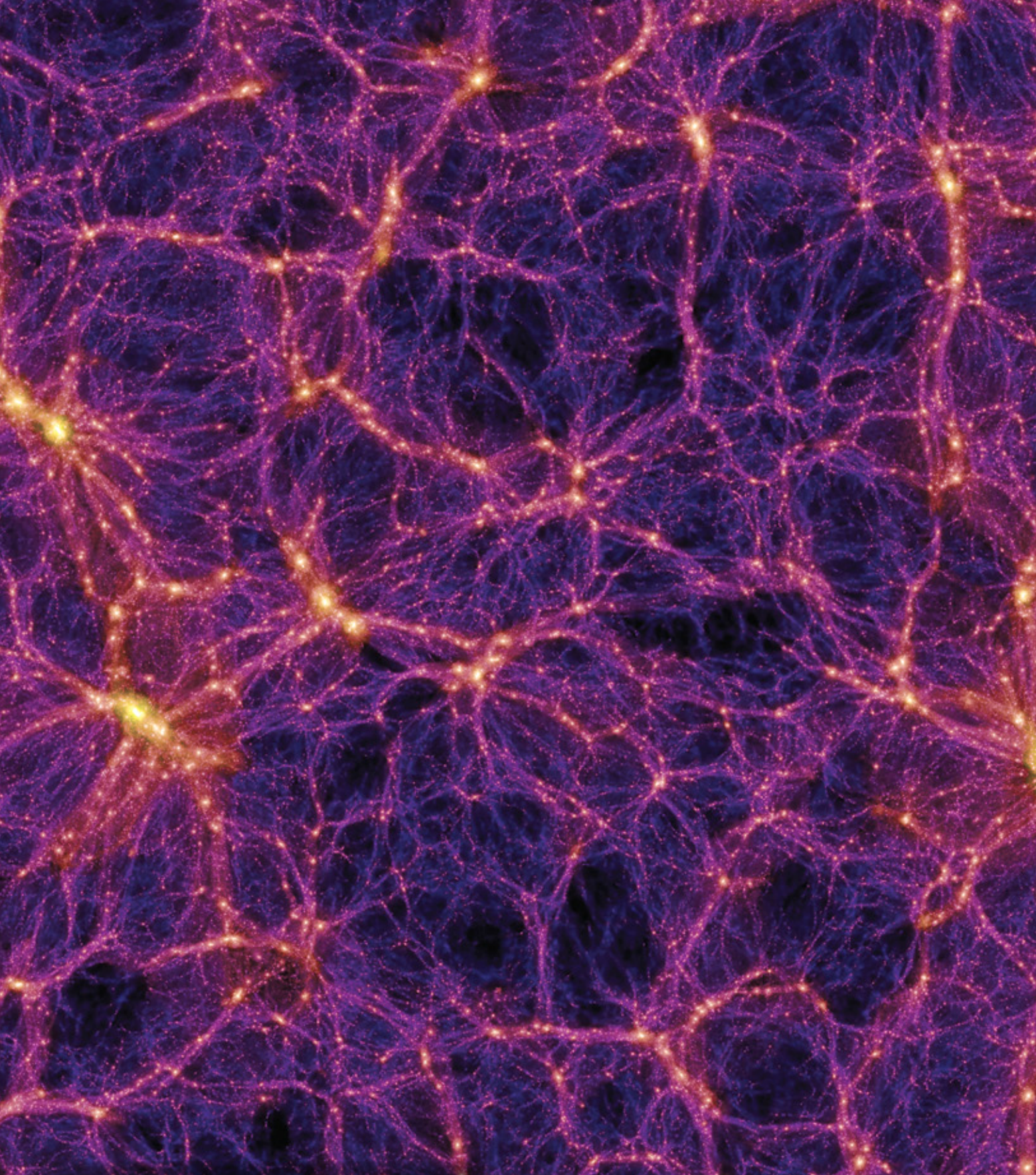
Formation of Supermassive Black Holes. Marta Volonteri in *Astronomy and Astrophysics Review,* Vol. 18, No. 3, pages 279–315; July 2010. http://arxiv.org/abs/1003.4404

Supermassive Black Holes Do Not Correlate with Galaxy Disks or Pseudobulges. John Kormendy, R. Bender and M. E. Cornell in *Nature,* Vol. 469, pages 374–376; January 20, 2011.

An Actively Accreting Massive Black Hole in the Dwarf Starburst Galaxy Henize 2-10. Amy E. Reines et al. in *Nature,* Vol. 470, pages 66–68; February 3, 2011.

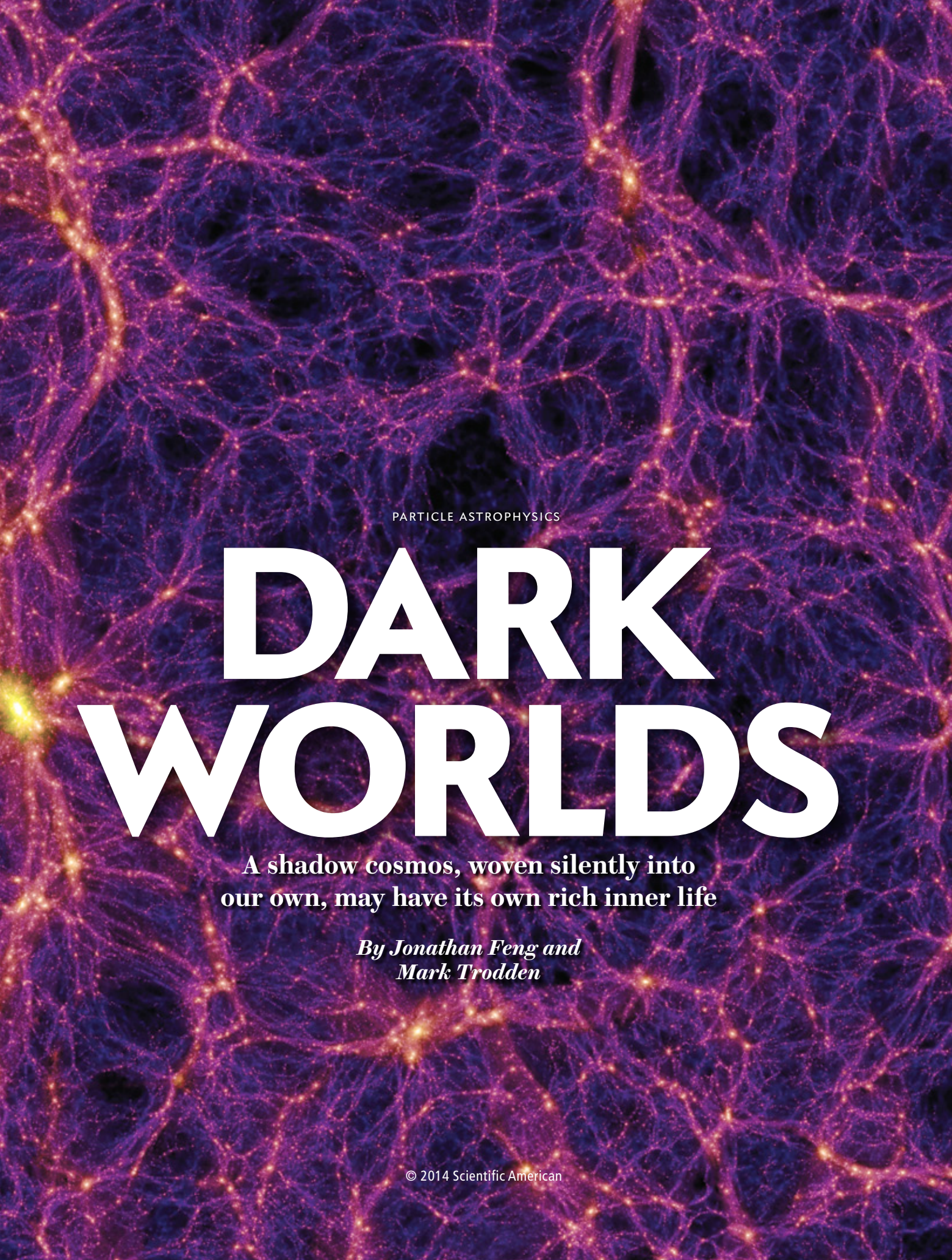
scientificamerican.com/magazine/sa

© 2014 Scientific American



THROUGH THE FORCE OF GRAVITY, dark matter sculpts the universe into a web of galaxies. Theorists now suspect that it may exert other forces as well. This image from the Millennium Simulation project in 2005 depicts a region roughly 1.6 billion light-years across.

© 2014 Scientific American



PARTICLE ASTROPHYSICS

DARK WORLDS

A shadow cosmos, woven silently into our own, may have its own rich inner life

By Jonathan Feng and Mark Trodden

© 2014 Scientific American

ON SEPTEMBER 23, 1846, JOHANN GOTTFRIED GALLE, DIRECTOR OF THE Berlin Observatory, received a letter that would change the course of astronomical history. It came from a Frenchman, Urbain Le Verrier, who had been studying the motion of Uranus and concluded that its path could not be explained by the known gravitational forces acting on it. Le Verrier suggested the existence of a hitherto unobserved object whose gravitational pull was perturbing Uranus's orbit in precisely the way required to account for the anomalous observations. Following Le Verrier's directions, Galle went to his telescope that night and discovered the planet Neptune.

A similar drama—in which astronomers observe anomalous cosmic motions, deduce the presence of new matter and go out to hunt for it—is playing out again today in modern cosmology. In the role of Uranus, we see stars and galaxies moving in ways they should not; in the role of Neptune, we deduce the existence of hitherto unobserved substances, provisionally called dark matter and dark energy. From the types of anomalies we see, we can glean a few basic facts about them. Dark matter seems to be a sea of invisible particles that fills space unevenly; dark energy is spread out uniformly and acts as if it is woven into the fabric of space itself. Scientists have yet to repeat Galle's accomplishment of pointing an instrument at the sky and glimpsing the unseen players definitively, but tantalizing inklings, such as blips in particle detectors, continue to accumulate.

From its discovery as a shadowy force on Uranus, Neptune proved to be a fascinating world in its own right. Might the same be true of dark matter and dark energy? Scientists are increasingly considering the possibility that dark matter, in particular, is not just a contrivance to account for the motion of visible matter but a hidden side of the universe with a rich inner life. It may consist of a veritable zoo of particles interacting through novel forces of nature—an entire universe interwoven silently with our own.

THE DARK SIDE

THESE IDEAS ARE A SHIFT from the long-held assumption that dark matter and dark energy are the most antisocial substances in the cosmos. Since astronomers first inferred the existence of dark matter in the 1930s, they have considered inertness its defining property. Observations suggest it outweighs ordinary matter by a factor of 6 to 1. Galaxies and galaxy clusters are embedded in giant balls, or "halos," of dark matter. For such a mass of material to elude direct detection, astronomers reason that it has to consist of particles that scarcely interact with ordinary matter or, indeed, with one another. All they do is provide the gravitational scaffolding for luminous matter.

Astronomers think the halos formed early on in cosmic history and then drew in ordinary matter, which, being capable of a rich range of behaviors, developed into intricate structures, while dark matter, being inert, remained in its primitive state. As for dark energy, its only role appears to be to accelerate cosmic expansion, and the available evidence indicates it has remained completely unchanged over the life of the cosmos.

The prospect that dark matter might be rather more interesting is driven not so much by the field of astronomy but by detailed investigations of the inner workings of atoms and the world of subatomic particles. Particle physicists have a tradition of seeing glimmers of unknown forms of matter in the behavior of known matter, and their evidence is completely independent of cosmic motions.

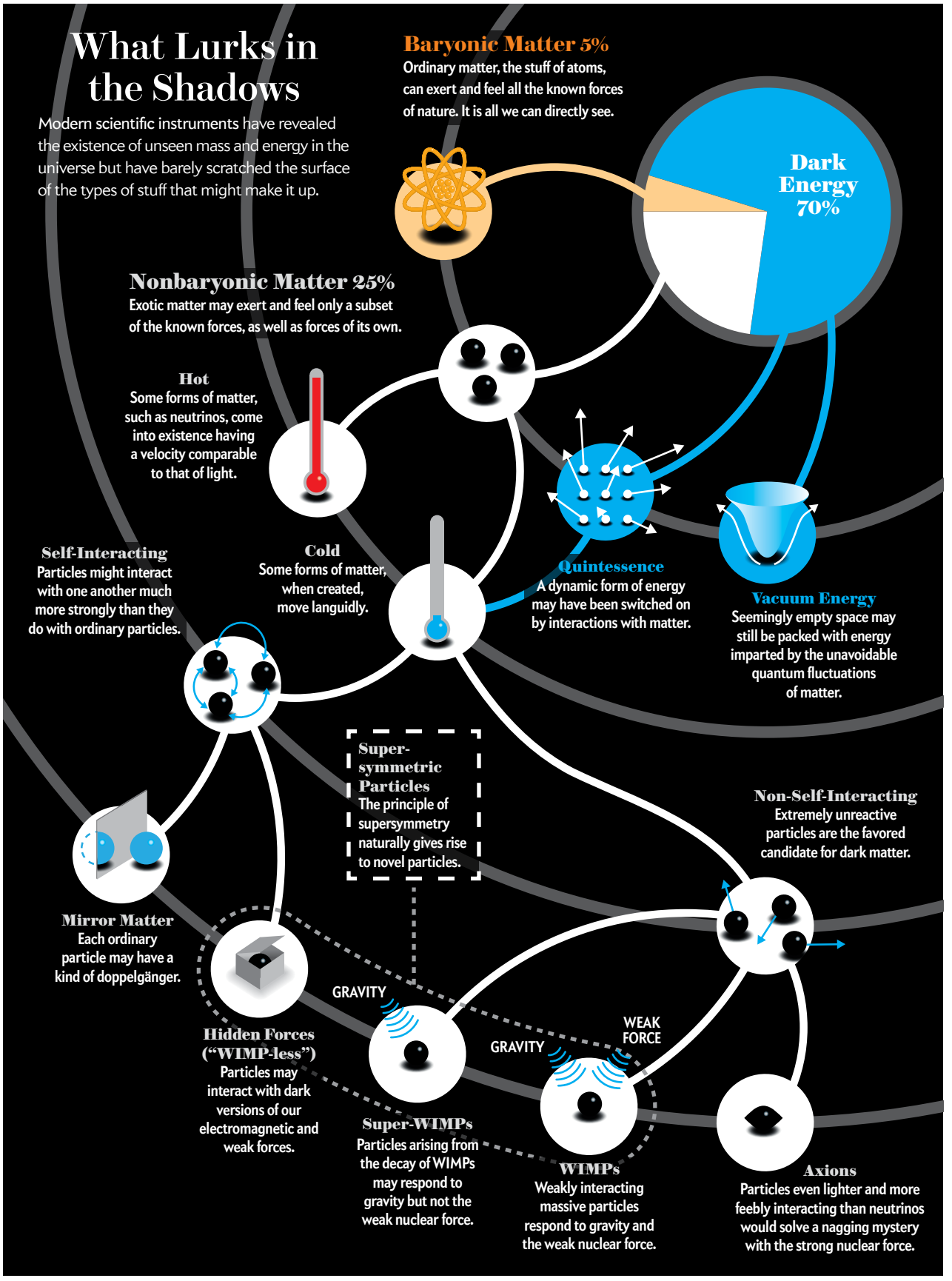
In the case of dark matter, the train of thought began with the discovery of radioactive beta decay in the early 1900s. Italian theorist Enrico Fermi sought to explain the phenomenon by postulating a new force of nature and new force-carrying particles that caused atomic nuclei to decay. This new force was similar to electromagnetism and the new particles to photons, the particles of light—but with a key twist. Unlike photons,

IN BRIEF

Scientists have two independent reasons for thinking that the cosmos is filled with some unknown form of matter, dark matter. Not only do stars, galaxies and gas clouds move as if they are being tugged by the gravity of hidden material, but processes such as radioactivity present puzzles that can be solved by the existence of hitherto unknown particles.
Dark matter is usually assumed to consist of WIMPs, a kind of particle that scarcely interacts with the visible world. Boringness is its sine qua non.
Or at least that is the usual assumption. Might dark matter in fact have a rich inner life? Particle physicists striving to understand what makes up dark matter think it could interact through a full range of forces, including a form of light to which our eyes are totally blind.

PRECEDING PAGES: COURTESY OF VOLKER SPRINGEL *Heidelberg University and Heidelberg Institute for Theoretical Studies*

© 2014 Scientific American


What Lurks in the Shadows
Modern scientific instruments have revealed the existence of unseen mass and energy in the universe but have barely scratched the surface of the types of stuff that might make it up.
Baryonic Matter 5%
Ordinary matter, the stuff of atoms, can exert and feel all the known forces of nature. It is all we can directly see.
Dark Energy 70%
Nonbaryonic Matter 25%
Exotic matter may exert and feel only a subset of the known forces, as well as forces of its own.
Hot
Some forms of matter, such as neutrinos, come into existence having a velocity comparable to that of light.
Cold
Some forms of matter, when created, move languidly.
Self-Interacting
Particles might interact with one another much more strongly than they do with ordinary particles.
Quintessence
A dynamic form of energy may have been switched on by interactions with matter.
Vacuum Energy
Seemingly empty space may still be packed with energy imparted by the unavoidable quantum fluctuations of matter.
Super-symmetric Particles
The principle of supersymmetry naturally gives rise to novel particles.
Non-Self-Interacting
Extremely unreactive particles are the favored candidate for dark matter.
Mirror Matter
Each ordinary particle may have a kind of doppelgänger.
GRAVITY
GRAVITY
WEAK FORCE
Hidden Forces ("WIMP-less")
Particles may interact with dark versions of our electromagnetic and weak forces.
Super-WIMPs
Particles arising from the decay of WIMPs may respond to gravity but not the weak nuclear force.
WIMPs
Weakly interacting massive particles respond to gravity and the weak nuclear force.
Axions
Particles even lighter and more feebly interacting than neutrinos would solve a nagging mystery with the strong nuclear force.

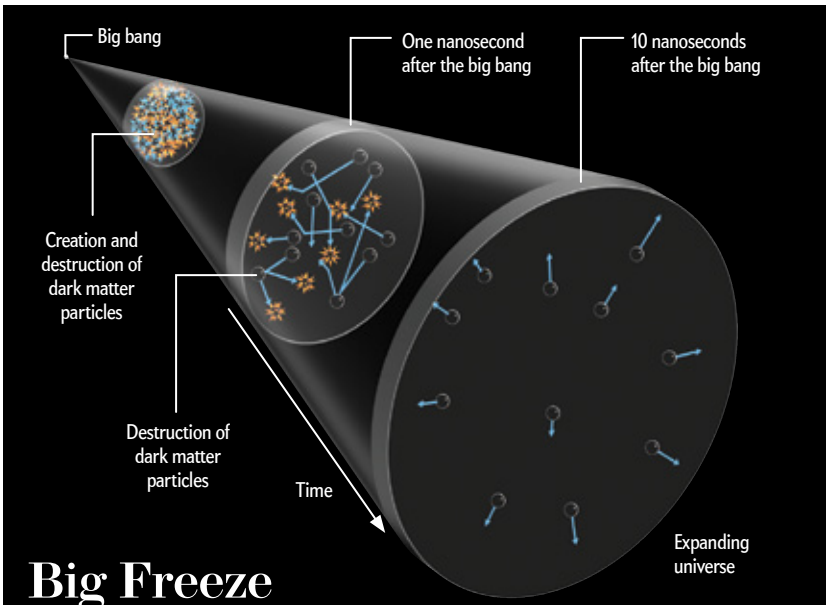
Illustrations by Bryan Christie

© 2014 Scientific American

which are massless and therefore highly mobile, Fermi argued that the new particles had to be heavy. Their mass would limit their range and account for why the force causes nuclei to fall apart but otherwise goes unnoticed. To reproduce the observed half-life of radioactive isotopes, they had to be quite heavy—around 100 times that of the proton, or about 100 giga-electron-volts, in the standard units of particle physics.

The new force is now known as the weak nuclear force, and the hypothesized force-carrying particles are the *W* and *Z* particles, which were discovered in the 1980s. They are not dark matter themselves, but their properties hint at dark matter. A priori, they should not be so heavy. Their high mass suggests that something is acting on them—novel particles that cause them to take on mass like a friend who encourages you to give into temptation and eat another slice of cake. One goal of the Large Hadron Collider is to look for those particles, which should have masses comparable to those of the *W* and *Z*. Indeed, physicists think dozens of types of particles may be waiting to be discovered—one for each of the known particles, paired off in an arrangement known as supersymmetry.

WHERE DARK MATTER CAME FROM


Big bang
One nanosecond after the big bang
10 nanoseconds after the big bang
Creation and destruction of dark matter particles
Destruction of dark matter particles
Time
Expanding universe

Big Freeze

In the hot, dense early universe, dark matter particles such as WIMPs were created and destroyed in a dynamic equilibrium. As the cosmos expanded, it cooled and eventually was no longer able to create new particles. Those left over became so spread out that they ceased colliding and getting destroyed. For WIMPs, theory makes a firm prediction for the amount of material that survived, which is consistent with observations.

These hypothetical particles include some collectively known as weakly interacting massive particles, or WIMPs. The name arises because the particles interact only by means of the weak nuclear force. Being immune to the electric and magnetic forces that dominate the everyday world, they are totally invisible and have scarcely any direct effect on normal particles. Therefore, they make the perfect candidate for cosmic dark matter.

Whether they can truly explain dark matter, though, depends on how many of them there are. Here is where the particle physics argument really gains traction. Like any other breed of particle, WIMPs would have been produced in the fury of the big bang. High-energy particle collisions back then both created and destroyed WIMPs, allowing a certain number of them to exist at any given moment. This number varied with time depending on two competing effects driven by the expansion of the universe. The first was the cooling of the primordial soup, which reduced the amount of energy available to create WIMPs, so that their number diminished. The second effect was the dilution of particles, which reduced the frequency of collisions until they effectively ceased to occur. At that point, about 10 nanoseconds after the big bang, the number of WIMPs became frozen in. The universe no longer had either the energy needed to create WIMPs or the dense concentrations of mass needed to destroy them.

Given the expected mass of WIMPs and the strength of their interactions, which govern how often they annihilate one another, physicists can easily calculate how many WIMPs should be left over. Rather amazingly, the number matches the number required to account for cosmic dark matter today, within the precision of the mass and interaction-strength estimates. This remarkable agreement is known as the WIMP coincidence. Thus, particles motivated by a century-old puzzle in particle physics beautifully explain cosmological observations.

This line of evidence, too, indicates that WIMPs are inert. A quick calculation predicts that nearly one billion of these particles have passed through your body since you started reading this article, and unless you are extraordinarily lucky, none has had any discernible effect. Over the course of a year you might expect just one of the WIMPs to scatter off the atomic nuclei in your cells and deposit some meager amount of energy. To have any hope of detecting such events, physicists set their particle detectors to monitor large volumes of liquid or other material for long periods. Astronomers also look for bursts of radiation in the galaxy that mark the rare collision and annihilation of orbiting WIMPs. A third way to find WIMPs is to try to synthesize them in terrestrial experiments [*see box on page 50*].

OUT-WIMPING THE WIMPS

THE EXTRAORDINARY EFFORT now being devoted to WIMP searches might leave the impression that these particles are the only theoretically plausible dark matter candidate. Are they? In fact, recent developments in particle physics have uncovered other possibilities. This work hints that the WIMP is just the tip of the iceberg. Lurking under the surface could be hidden worlds, complete with their own matter particles and forces.

One such development is the concept of particles even more wimpy than WIMPs. Theory suggests that WIMPs formed in the first nanosecond of cosmic history might have been unstable. Seconds to days later they could have decayed to particles that have a comparable mass but do not interact by the weak nucle-

© 2014 Scientific American

ar force; gravity is their only connection to the rest of the natural world. Physicists, tongue in cheek, call them super-WIMPs.

The idea is that these particles, rather than WIMPs, constitute the dark matter of today's universe. Super-WIMPs would elude direct observational searches but might be inferred from the telltale imprint they would leave on the shapes of galaxies. When created, super-WIMPs would have been moving at a significant fraction of the speed of light. They would have taken time to come to rest, and galaxies could not have begun forming until they did. This delay would have left less time for matter to accrete onto the centers of galaxies before cosmic expansion diluted it. The density at the center of dark matter halos should therefore reveal whether they are made of WIMPs or super-WIMPs; astronomers are now checking. In addition, the decay from WIMP to super-WIMP should have produced photons or electrons as a by-product, and these particles can smash into light nuclei and break them apart. There is some evidence that the universe has less lithium than expected, and the super-WIMP hypothesis is one way to explain the discrepancy.

The super-WIMP scenario also inspires fresh possibilities for what experimental physicists might observe. For instance, the original WIMP need not have been either dark or wimpy; it could have had an electric charge. Any charge it had would not have affected the evolution of the cosmos, because the particle decayed so quickly. It would, however, mean that WIMPs would be extremely conspicuous if experimentalists were able to re-create them. Particle detectors would register them as electrons on steroids; having the same charge as an electron but 100,000 times more mass, such a particle would barrel through the detectors, leaving spectacular tracks in its path.

DARK FORCES, HIDDEN WORLDS

THE MAIN LESSON of super-WIMP models is that there is no reason, either theoretically or observationally, that dark matter should be as boring as astronomers tend to presume. Once one admits the possibility of hidden particles with properties that go beyond the standard WIMP scenario, it is natural to consider the full range of possibilities. Could there be a whole sector of hidden particles? Could there be a hidden world that is an exact copy of ours, containing hidden versions of electrons and protons, which combine to form hidden atoms and molecules, which combine to form hidden planets, hidden stars and even hidden people?

The possibility that a hidden world could be identical to ours has been explored at length, beginning in 1956 with an offhand comment in a Nobel Prize–winning paper by Tsung-Dao Lee and Chen Ning Yang and more recently by many others, including Robert Foot and Raymond Volkas of the University of Melbourne in Australia. The idea is truly tantalizing. Could it be that what we see as dark matter is really evidence for a hidden world that mirrors ours? And are hidden physicists and astronomers even now peering through their telescopes and wondering what their dark matter is, when in fact their dark matter is us?

Unfortunately, basic observations indicate that hidden worlds cannot be an exact copy of our visible world. For one, dark matter is six times more abundant than normal matter. For another, if dark matter behaved like ordinary matter, halos would have flattened out to form disks like that of the Milky Way—with dramatic gravitational consequences that have not been seen. Last, the existence of hidden particles identical to ours would have affect-

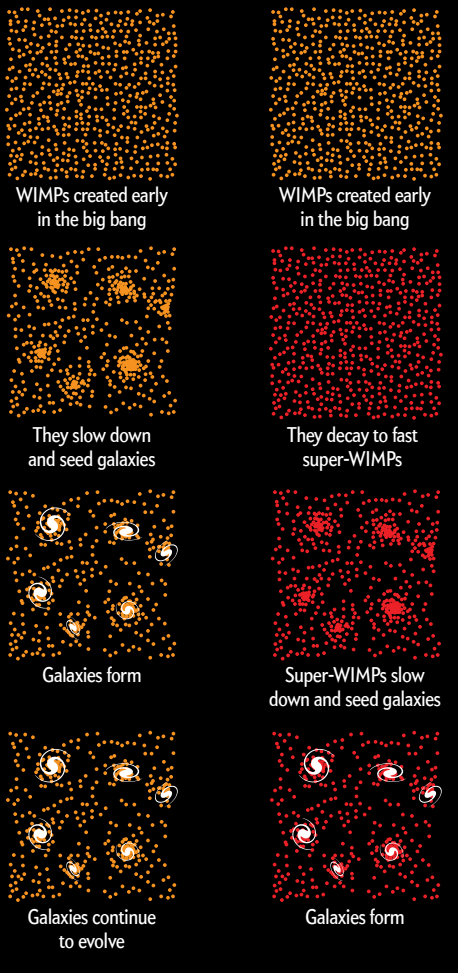
NEW TYPE OF DARK MATTER

Varieties of Weaklings

Super-WIMPs were the first proposed type of particle that enriches the standard WIMP scenario for dark matter. The term is intentionally ironic: these particles are "super" not because they are mightier than WIMPs but because they are wimpier: they interact with ordinary matter only through the force of gravity.

	Baryonic	WIMP	Super-WIMP
Gravity	✓	✓	✓
Electromagnetic force	✓		
Weak nuclear force	✓	✓	
Strong nuclear force	✓		
Possible dark forces		✓	✓

In the WIMP scenario (*left column*), WIMPs seed galaxy formation directly. In the super-WIMP scenario (*right column*), they decay to super-WIMPs, which do the seeding—with a delay relative to the WIMP scenario.


WIMPs created early in the big bang
WIMPs created early in the big bang
They slow down and seed galaxies
They decay to fast super-WIMPs
Galaxies form
Super-WIMPs slow down and seed galaxies
Galaxies continue to evolve
Galaxies form

© 2014 Scientific American

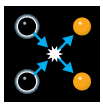
ed cosmic expansion, altering the synthesis of hydrogen and helium in the early universe; compositional measurements rule that out. These considerations argue strongly against hidden people.

That said, the dark world might indeed be a complicated web of particles and forces. Several investigators, including one of us (Feng) and Jason Kumar of the University of Hawaii at Manoa, have found that the same supersymmetric framework that leads to WIMPs allows for alternative scenarios that lack WIMPs but have multiple other types of particles. What is more, in many of these WIMP-less theories, these particles interact with one another through newly postulated dark forces. We found that such forces would alter the rate of particle creation and annihilation in the early universe, but again the numbers work out so that the right number of particles are left over to account for

PHYSICS OBSERVATIONS

How to See the Unseeable

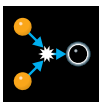
So far everything astronomers know about dark matter comes from its gravitational effects on visible matter. But they need to detect it directly if they are to find out what it is. That will not be easy: dark matter is elusive by definition. Nevertheless, motivated by the promise of discovering what a quarter of the universe is, thousands of researchers are looking. Most of their efforts have focused on WIMPs, and the three common search strategies are to look for the particles' annihilation, scattering and production.



ANNIHILATION When two WIMPs meet, they obliterate each other and leave behind a clutch of other particles such as electrons, antielectrons (known as positrons) and neutrinos. Such annihilation cannot be very common, or else no WIMPs would be left by now. Fortunately, current experiments are sensitive enough to notice if even a minuscule fraction of WIMPs are being annihilated.

Detectors on high-altitude balloons and satellites have sought electrons and positrons. The Alpha Magnetic Spectrometer is mounted to the International Space Station in low-Earth orbit, looking for positrons from dark matter annihilation. Other observatories such as the Super-Kamiokande experiment in Japan and IceCube in Antarctica are also watching for neutrinos.

DIRECT DETECTION Dark matter should be streaming through our planet as it travels through the galaxy. On rare occasions, a WIMP will bump into an atomic nucleus and cause it to recoil, just as a pool ball does when struck by the cue ball. The predicted recoil energies are almost imperceptible but may be within the range of sensitive detectors. Cryogenic technology slows the natural vibrations of atoms and makes it easier to notice recoil. The energy deposited in the detector holds the key to pinning down the fundamental properties of dark matter. Two experiments, DAMA and CoGeNT, have claimed to detect a signal (*below*), but others, such as XENON and CDMS, have found nothing of significance. These and other new experiments are improving their sensitivities rapidly, promising an exciting future for this field.



PRODUCTION Dark matter might be created at particle colliders, such as the Large Hadron Collider at CERN near Geneva, a mammoth experiment that collides protons together at extremely high energies. Dark matter production is dark matter annihilation played backward: if dark matter can annihilate into normal particles, it can also be produced by the collisions of normal particles. The signature of dark matter production would be the observation of collisions in which energy and momentum seem to go missing, indicating that some unreactive particles have been produced and then escaped the detector without registering. These giant experiments, designed to tease out the secrets of the subatomic world, may wind up discovering the dominant form of matter in the universe.

Experiments That Claim to Have Detected Dark Matter Particles

EXPERIMENT	CDMS	DAMA	CoGeNT	PAMELA
What it stands for	Cryogenic Dark Matter Search	DArk MAtter	Coherent Germanium Neutrino Technology	Payload for Antimatter Matter Exploration and Light-nuclei Astrophysics
Where it is	Soudan mine in Minnesota	Gran Sasso underground lab in Italy	Soudan mine	Attached to Russian satellite
What it has seen	Recoil events	Annual variation in the number of recoil events	Recoil events	Excess of positrons
Why the signal might be real	Direct, expected signal of dark matter	Statistically significant	Sensitive to ultralow-energy recoils	Direct, expected signal of dark matter annihilation
Why it might not be	Not statistically significant	Apparently excluded by other results	Normal nuclear events might be responsible	Could be explained by astrophysical sources
Which experiments are following up	SuperCDMS, XENON, LUX	XENON, MAJORANA Demonstrator, LUX	XENON, MAJORANA Demonstrator, LUX	Alpha Magnetic Spectrometer

© 2014 Scientific American

dark matter. These models predict that dark matter may be accompanied by a hidden weak force or, even more remarkably, a hidden version of electromagnetism, implying that dark matter may emit and reflect hidden light.

This "light" is invisible to us, and so the dark matter remains unseen. Still, new forces could have very significant effects. For example, they could distort interacting clouds of dark particles. Astronomers have sought this effect in the famous Bullet cluster, which consists of two clusters of galaxies that have passed through each other. Observations show the co-mingling of clusters left the dark matter largely unperturbed, indicating that any dark forces are weak. Researchers are continuing to look in other systems.

Such forces would also let dark particles exchange energy and momentum with one another, a process that would tend to homogenize them and cause once lopsided halos to become spherical. This process should be most pronounced for dwarf galaxies, where the dark matter is slow-moving, particles linger near one another and small effects have time to build up. The observation that small galaxies are systematically rounder than their larger cousins would suggest dark matter interacting through new forces. Astronomers are only just beginning these studies.

FROM ONE DARK THING TO ANOTHER

AN EQUALLY INTRIGUING possibility is that dark matter interacts with dark energy. Most theories treat the two as disconnected, but there is no reason they must be, and physicists are now considering how the two might affect each other. One hope is that couplings between them might mitigate some cosmological problems, such as the coincidence problem—the question of why the two have comparable densities. Dark energy is roughly three times as dense as dark matter, but the ratio might have been far larger. This coincidence would make sense if dark matter somehow triggered the emergence of dark energy.

Couplings with dark energy might also allow dark matter particles to interact in ways that ordinary particles do not. Recent models allow and sometimes even mandate dark energy to exert a different force on dark matter than it does on ordinary matter. Under the influence of this force, dark matter would tend to pull apart from any ordinary matter it had been interlaced with. In 2006 Marc Kamionkowski, then at the California Institute of Technology, and Michael Kesden, then at the Canadian Institute for Theoretical Astrophysics, suggested looking for this effect in dwarf galaxies being torn apart by their larger neighbors. The Sagittarius dwarf galaxy, for example, is being eaten by the Milky Way, and astronomers think its dark matter and ordinary matter are spilling into our galaxy. They calculate that if the forces acting on dark matter are at least 4 percent stronger or weaker than the forces acting on ordinary matter, then the two components should drift apart by an observable amount. At present, however, the data show nothing of the sort.

Another idea is that interacting dark matter and dark energy would alter the growth of cosmic structures, which depends delicately on the composition of the universe. A number of researchers, including one of us (Trodden), with collaborators Rachel Bean, Éanna Flanagan and Istvan Laszlo of Cornell University, have used this constraint to rule out a large class of models.

Despite these null results, the theoretical case for a complex dark world is now so compelling that many researchers would find it more surprising if dark matter turned out to be nothing more than an undifferentiated swarm of WIMPs. After all, visible matter comprises a rich spectrum of particles with multiple interactions determined by beautiful underlying symmetry principles, and nothing suggests that dark matter and dark energy should be any different. We may not encounter dark stars, planets or people, but just as we could hardly imagine the solar system without Neptune, Pluto and the swarm of objects that lie even farther out, one day we might not be able to conceive of a universe without an intricate and fascinating dark world. SA

Jonathan Feng is a professor at the University of California, Irvine, working at the intersection of particle physics and cosmology, with a focus on dark matter.
Mark Trodden is co-director of the Center for Particle Cosmology at the University of Pennsylvania and an author of *Cosmic Variance*, a leading physics blog.

ASTRONOMICAL OBSERVATIONS

Silver Bullet

The famous Bullet cluster is among astronomers' most persuasive evidence for dark matter. It is actually a pair of galaxy clusters that collided. The collision did not affect the galaxies' stars (*visible image*), because they present small targets on these scales, but interstellar gas clouds rammed into one another and emitted x-rays (*pink*). Dark matter (*blue*) betrayed its presence because its gravity distorted the light of background objects. It remained aligned with the stars—indicating that whatever particles make it up are highly unreactive.

NASA/CXC/M. MARKEVITCH ET AL. (*x-ray*); NASA/STSCI AND MAGELLAN/U. ARIZONA/D. CLOWE ET AL. (*optical*); NASA/STSCI, ESO, WFI AND MAGELLAN/U. ARIZONA/D. CLOWE ET AL. (*lensing map*)

MORE TO EXPLORE

Einstein's Telescope: The Hunt for Dark Matter and Dark Energy in the Universe. Evalyn Gates. W. W. Norton, 2009.
Approaches to Understanding Cosmic Acceleration. Alessandra Silvestri and Mark Trodden in *Reports on Progress in Physics*, Vol. 72, No. 9, Paper No. 096901; September 2009. arxiv.org/abs/0904. 0024
Modern Cosmology and the Building Blocks of the Universe. Mark Trodden. Penn Alumni Weekend Lecture, May 15, 2010. www.sas.upenn.edu/home/news/troddenandabbate/trodden.html
What's the Matter? The Search for Clues in Our Cold, Dark Universe. Jonathan Feng. Pagels Memorial Public Lecture, July 14, 2010. vod.grassrootstv.org/vodcontent/9251-1.wmv
Dark Matter Candidates from Particle Physics and Methods of Detection. Jonathan L. Feng in *Annual Reviews of Astronomy and Astrophysics*, Vol. 48, pages 495–545; August 2010. arxiv.org/abs/1003.0904

scientificamerican.com/magazine/sa

© 2014 Scientific American

GALAXY FORMATION

THE LOST GALAXIES

By the latest estimate, the observable universe contains 200 billion galaxies. Astronomers wonder: Why so few?

By James E. Geach

IN BRIEF

Forget dark matter: even the supposedly normal matter of the universe is mysterious enough. Why does only a small fraction of it reside in galaxies? Where did the rest go?

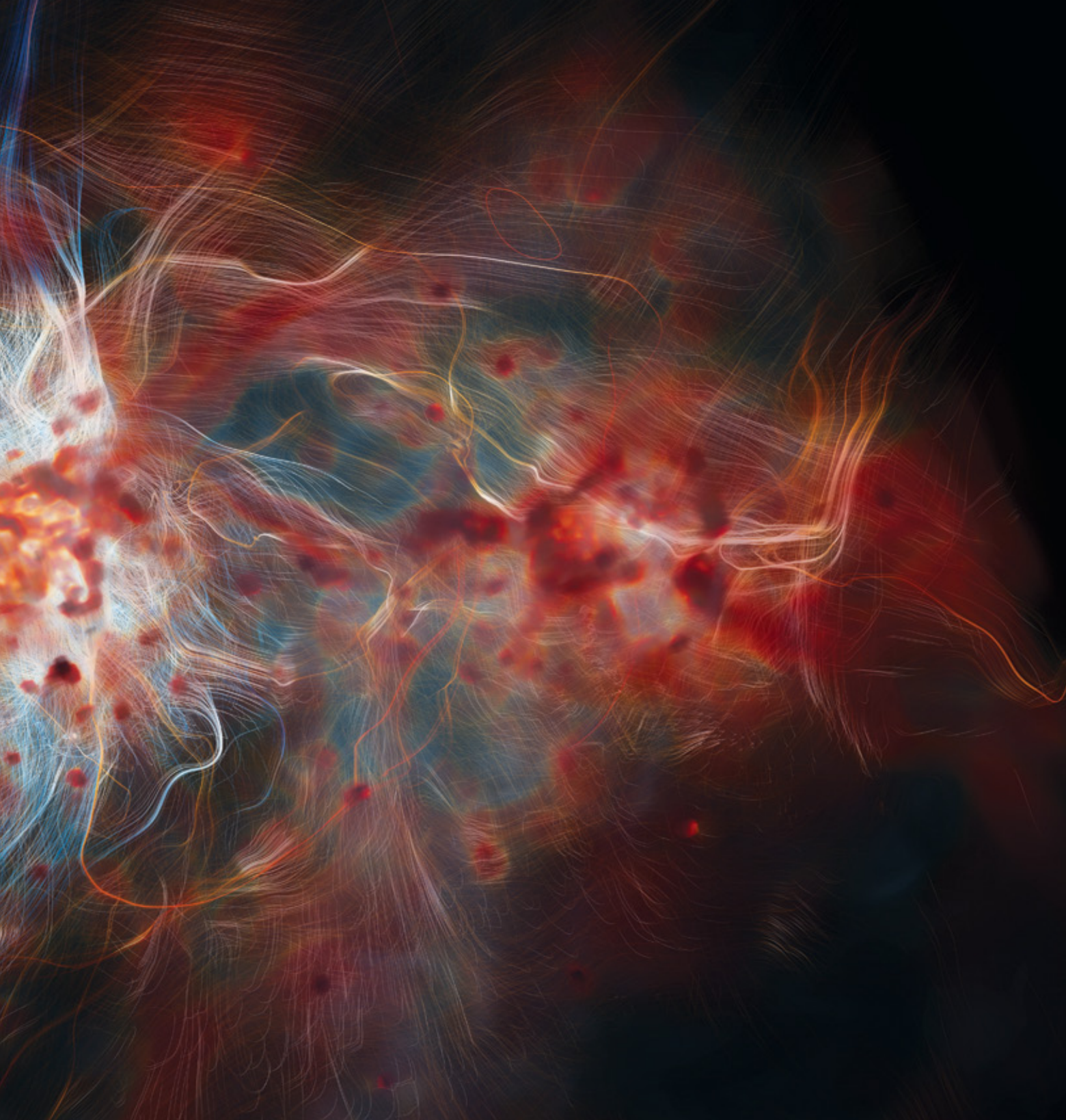
The current best guess is that the bulk of the normal matter is trapped in giant gaseous filaments. This so-called warm-hot intergalactic medium, or WHIM, is hard to detect directly.

Galaxy formation is evidently rather inefficient. As material falls into a galaxy, the galaxy tends to shoot much of it straight back out again—a process known as feedback.

The atoms in your body have probably been cycled through intergalactic space. Indeed, galaxies and their contents are not fixed structures but the bright tips of a wider sea of gas.

Graphic by James E. Geach and Rob Crain

© 2014 Scientific American



IN THE THROES OF FORMATION, a galaxy like our Milky Way pulls in cold, dense gas (*red streams*) and ejects hot gas (*blue streams*) back into intergalactic space. The galaxy ends up with only a small fraction of the raw material. The author and his colleagues generated this image using a state-of-the-art cosmological simulation code.

© 2014 Scientific American

I HAVE ALWAYS BEEN STARTLED AND FASCINATED BY THE SANDLIKE ABUNDANCE of galaxies sprinkled across the night sky. One of the most sensitive optical images ever made by human beings, the Hubble Ultra Deep Field, captures some 10,000 galaxies in an area about 1/100th the size of the full moon. Scaled up to the whole sky, such a density implies a total of 200 billion or so galaxies. And those are just the most luminous ones; the true number is probably much larger. How did all those galaxies come to be? This question inspired me to become an astronomer and has been the focus of my research career. Over the years my naive way of looking at galaxies has changed. To judge by their sheer numbers, nature appears to be quite good at producing galaxies. Not so. If you add up all the visible matter in galaxies today, you get only about a tenth of the total endowment created by the big bang. Where is the rest, and why did it not end up in galaxies? These are two of the biggest puzzles in astronomy today.

This missing matter is different from dark matter and dark energy. Those are substances of unknown composition that together amount to 96 percent of the total mass of the cosmos. In the case of numbers of galaxies, the trouble is with the 5 percent that was supposed to be well understood. This slice of the universe is normal matter, made of the same stuff as our bodies and everything around us—primarily baryons, the class of particle that includes protons and neutrons. For the majority of it to go missing is a mystery inside a mystery. Not only is most of the matter in the universe dark and unexplained, but of the small sliver that is normal, only a fraction is accounted for.

Another way to put it is that the process of galaxy formation must be inefficient. It is as if a farmer sowed an entire field of seeds and only one of every 10 germinated. Astronomers have struggled for years to explain how that could be. The emerging answer requires us to revise our notions not only of how galaxies form but of what a "galaxy" even is. One expects not to fathom exotic types of matter; it is rather more disturbing to learn that we still do not grasp the mundane sort.

HONEY, WE LOST HALF THE UNIVERSE

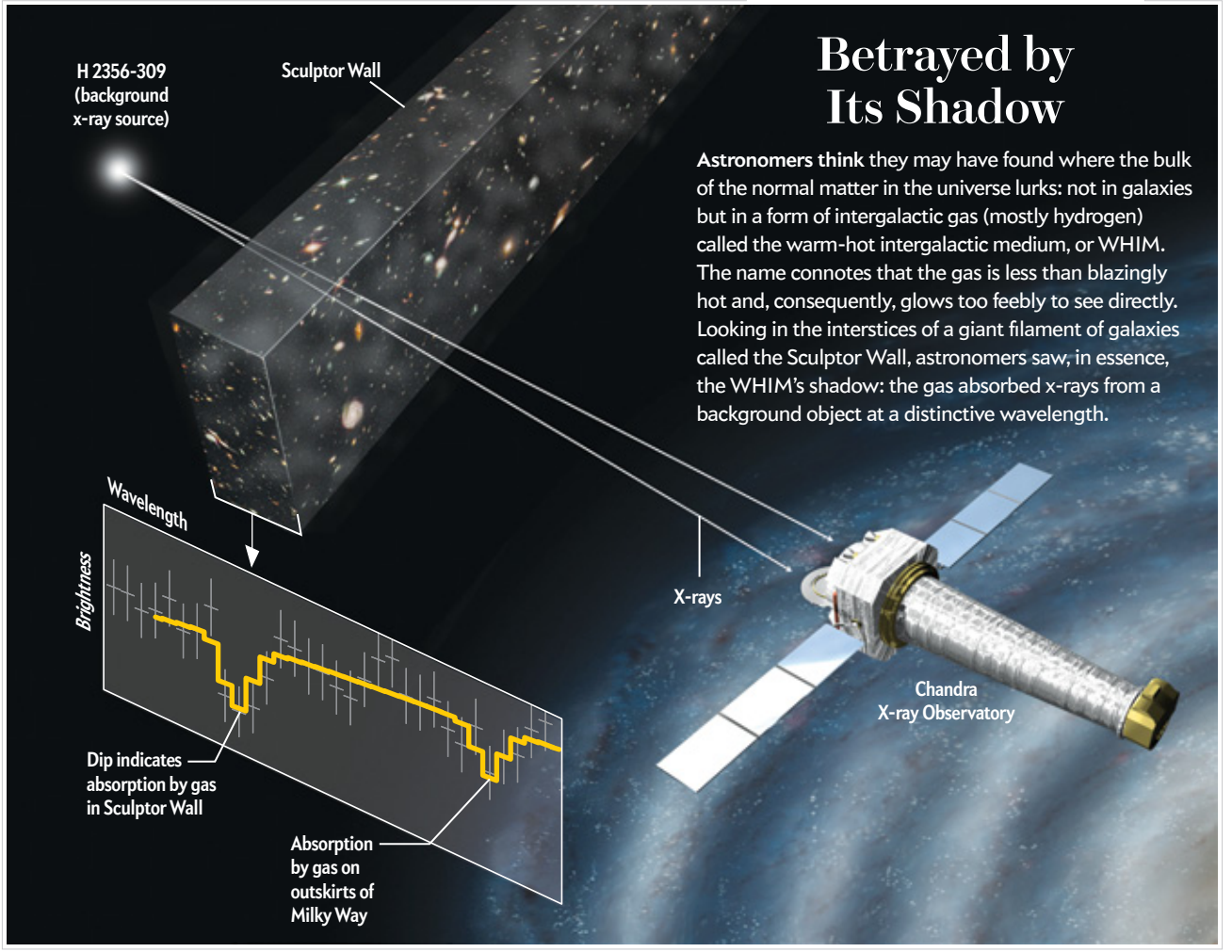
FOR DECADES observers have been piecing together a timeline of cosmic history that describes the content of the universe at different stages in its evolution. In the process, it has gradually become apparent that more baryons were present at the dawn of the universe than we can directly detect today.

The initial amount of baryonic matter is actually fairly straightforward to estimate. This information is encoded into the relic radiation from the big bang: the cosmic microwave background radiation. State-of-the-art experiments such as the Wilkinson Microwave Anisotropy Probe and the Planck space observatory detect small fluctuations in the temperature of this radiation, and the distribution of these fluctuations reflects the baryon density of the universe at a time when galaxies were yet to form. An independent check comes from measurements of the abundances of helium, deuterium and lithium. These elements were synthesized in the first few minutes of the universe in relative amounts that depended on the overall quantity of baryonic matter. Both techniques imply that the total amount of baryonic matter should account for about 5 percent of the mass of today's universe.

At the outset, all baryons took the form of a hot gas that filled space. In the regions where the initial matter density was high, gravity caused gas to clump into progressively denser clouds, which were the starting point for galaxy formation. Astronomers have detected this reservoir of gas in the early universe by analyzing the light detected from bright, distant quasars. What quasars are is not important for now; just consider them extremely bright lighthouses, acting as backlighting for the primordial gas floating around in intergalactic space. When a light ray from a quasar passes through a cloud of cold, neutral hydrogen, the gas absorbs some of the photons. Because the gas absorbs only photons of a certain energy, it imprints a telltale dip in the quasar's spectrum at a very specific wavelength: what astronomers call an absorption line.

A ray of light from the quasar might pass through hundreds of such clouds on its journey through the universe, and each one can imprint an absorption line at a slightly different wavelength, depending on the clouds' distance from the observer. By summing up the dips, we can calculate how many of the baryons were locked into these clouds. The result suggests that as late as five billion years after the big bang, or about nine billion years ago, the initial allotment of baryons could still be accounted for. Most were floating around in intergalactic space and had not yet collapsed into luminous galaxies [see "The Emptiest Places," by Evan Scannapieco, Patrick Petitjean and Tom Broadhurst; SCIENTIFIC AMERICAN, October 2002].

© 2014 Scientific American



Betrayed by Its Shadow

Astronomers think they may have found where the bulk of the normal matter in the universe lurks: not in galaxies but in a form of intergalactic gas (mostly hydrogen) called the warm-hot intergalactic medium, or WHIM. The name connotes that the gas is less than blazingly hot and, consequently, glows too feebly to see directly. Looking in the interstices of a giant filament of galaxies called the Sculptor Wall, astronomers saw, in essence, the WHIM's shadow: the gas absorbed x-rays from a background object at a distinctive wavelength.

In the intervening nine billion years, most of the galaxies we see today formed, wrought from that vast reservoir of primordial hydrogen. Once inside galaxies, baryons were reprocessed and took on various guises: stars, stellar remnants, neutral gas (both atomic and molecular), ionized gas, dust, planets, people. We can audit the mass of baryons in these different forms by measuring their emissions across the electromagnetic spectrum. For example, visible and near-infrared light reveal the mass of stars; a distinctive radio emission line signals the amount of neutral atomic hydrogen; infrared light betrays interstellar dust. In these ways, astronomers have taken a census of all the different phases of baryons in all the galaxies around us, and here is where the discrepancy arises: the total accounts for only 10 percent of the initial inventory of baryons that were present in the early universe. Presumably they did not simply vanish; they linger in the vast spaces between the galaxies. But why can we not see them?

ON A WHIM

ASTRONOMERS KNOW WHERE to find some of these intergalactic baryons. Dense swarms of galaxies called clusters are filled with a diffuse ionized gas, or plasma. The intense gravitational field of the cluster whips ions to high speed, giving the plasma a temperature of hundreds of millions of kelvins, enough to make it glow with x-rays. Space telescopes such as XMM-Newton and Chandra routinely detect clusters of galaxies by means of this x-ray emission. But clusters are rare, so the gas within them accounts for only another 4 percent of the baryons. When we add up all the baryons we can see in galaxies, clusters and elsewhere in intergalactic space, they account for roughly half the total, leaving the equivalent of at least 500 billion galaxies waiting to be found.

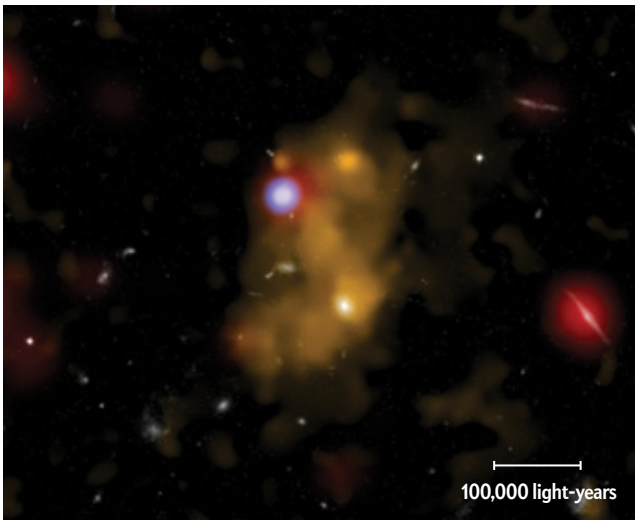
To balance the books, Renyue Cen of Princeton University, Jeremiah P. Ostriker, now at Columbia University, Romeel Davé of the University of Arizona and their collaborators conjectured a decade ago that the missing baryons are out there but have evolved into a phase that is hard to detect. The properties of this elusive component are related to the way that cosmic matter of all types, both mundane and exotic, has developed what astronomers refer to as large-scale structure.

Under the influence of gravity, dark matter has pulled itself into a vast skeletal network that interlaces the universe. Clusters are actually just the high-density nodes of this cosmic web. Outside clusters the majority of galaxies congregate in lower-density groups or line up in long filaments. Intergalactic gas is gravitationally attracted to filaments, and as it falls in, simulations sug-

SOURCE: "CONFIRMATION OF X-RAY ABSORPTION BY WARM-HOT INTERGALACTIC MEDIUM IN THE SCULPTOR WALL" BY TAOTAO FANG ET AL., IN *ASTROPHYSICAL JOURNAL*, VOL. 714, NO. 2, APRIL 22, 2010 (*graph*)

Illustration by Don Foley

© 2014 Scientific American



BEWARE OF THE BLOB: A wad of hydrogen gas (*yellow*) appears to be the castoff from the formation of a massive galaxy.

gest, it gets heated by shock waves to temperatures ranging from 100,000 kelvins to tens of millions of kelvins. That sounds hot but is tepid by the standards of intracluster gas. It is toasty enough to remain highly ionized but too cool to blaze in x-rays.

Cen, Ostriker and Davé dubbed this material the warm-hot intergalactic medium, or WHIM. If we could empirically confirm its presence and extent, we might be able to pin down the location and condition of the missing baryons.

The most promising way to detect the WHIM is to look for trace constituents such as ionized oxygen or nitrogen, which absorb ultraviolet light or x-rays of distinctive wavelengths. In fact, astronomers can apply the same absorption-line technique we used for the census of cold hydrogen clouds in the early universe; namely, we can look for dips in the spectra of quasars that backlight the WHIM. We have already had tantalizing glimpses. In the ultraviolet, the Hubble Space Telescope and the now defunct Far Ultraviolet Spectroscopic Explorer (FUSE) have detected absorption by strongly ionized oxygen. The first hints came over a decade ago, when the concept of the WHIM was still novel. Todd M. Tripp, now at the University of Massachusetts Amherst, and Blair D. Savage of the University of Wisconsin–Madison detected ionized oxygen absorption in the far-ultraviolet spectrum of the quasar PG 0953+415. More observations have followed over the past decade, helped by improvements in detector technology and instrumentation, the most recent being the installation of the Cosmic Origins Spectrograph on Hubble. Although systems with strongly ionized oxygen appear to be plentiful, this ion traces only the relatively cool part of the WHIM. To trace the more abundant hotter gas, we must search for absorption by even more highly ionized species.

Taotao Fang of the University of California, Irvine, and his collaborators have used the x-ray telescopes Chandra and XMM-Newton to peek into the interstices of the Sculptor Wall, a vast string of galaxies in the local universe—perfect WHIM-hunting territory. They have found absorption by oxygen that is so strongly ionized that it has lost almost all its electrons. The team estimated that the overall baryon density in this WHIM component agreed with cosmological simulations.

Though encouraging, these observations only scratch the surface. The observations are hard: the WHIM signal is weak, and we are generally working at the technical limits of the instrumentation. Even when we do detect absorption, we have to make many assumptions about the makeup of the gas to extrapolate the wider WHIM properties. More important, the absorption-line technique relies on fortuitously placed quasars. Quasars are rare, bright ones even more so, which makes WHIM hunting somewhat of a lottery. Nevertheless, we think we know where the missing baryons are and how they might be detected. Many astronomers are now engaged in efforts to map out the WHIM properly.

THE BATTLE FOR BARYONS

THE EXISTENCE OF THE WHIM goes some way to explaining why galaxy formation is so inefficient. The evolution of large-scale structure made intergalactic gas too tenuous and hot to accumulate into the cool, dense pools required for galaxy formation. Obviously, though, some of the baryons did manage to turn into galaxies, or else we would not be here.

Another thing is clear, too: galaxy formation used to be much more efficient. About eight billion years ago the average birth rate of stars was 10 to 20 times higher than it is today. Most of the galaxies we see today took shape then. To account for why galaxy formation slackened as sharply as it has, astronomers have had to rethink our basic models of how galaxies are born.

In principle, the recipe for a galaxy is quite simple. In a model pioneered in the 1990s by Simon D. M. White of the Max Planck Institute for Astrophysics in Garching, Germany, and Carlos S. Frenk of Durham University in England, galaxies grow within massive clumps of dark matter, termed halos, whose gravitational attraction sucks in surrounding gas like water going down a plug hole. In this model, some of the gas gets heated by shock waves as it plows into the halo and then cools by emitting radiation, allowing it to agglomerate into a cohesive body. Once within the galaxy, the gas can cool further and collapse into clouds of molecular hydrogen. Under gravitational contraction, these clouds can eventually reach the density required to make stars. Bigger galaxies can grow through the mergers of smaller ones.

White and Frenk recognized that their model could not be the entire story, however. For instance, not all of the gas flowing into galaxies would be shock-heated to high temperatures. But the basic picture of gas accretion within dark halos gave astronomers a solid framework within which to understand the principles of galaxy formation. The field has blossomed in the past 20 years. Theorists have explored the physics of gas flow in ever more detail, refining the original model. Recent high-resolution computer models of the thermodynamic evolution of gas in cosmological simulations suggests that some of the gas flowing into young galactic disks in the early universe does so in streams that are relatively cold (10,000 to 100,000 kelvins) and narrow (a few thousand light-years across). These cold flows appear to penetrate the hotter halo gas and directly feed the galaxies.

No one has yet seen this process in action: the detailed physics of the accretion of gas onto galaxies is complicated, and different simulations predict slightly different things. These caveats aside, astronomers now accept that all galaxies build up from the gravitational accumulation of primordial gas, be it gas that heats up and cools down or gas that never heats up at all.

The trouble with this model is that the flow of gas into galax-

COURTESY OF NASA/CXC/J. GEACH ET AL./M. WEISS

© 2014 Scientific American

ies cannot go on unabated. If it did, galaxies would grow into monsters, and we know they do not: galaxies today come in only a limited range of masses. Early models seemed to reproduce the observed range of galactic masses pretty well, but in retrospect they worked only because astronomers were using a value for the overall baryon density that was about half the present value. As new measurements of the baryon fraction revised the value upward, theorists fed this information into simulations and realized that their model universes were plagued by a serious overabundance of massive galaxies that are not seen in nature.

Another problem is that models predict a profusion of smallish dark matter clumps that agglomerate into progressively larger bodies. Real galaxies do not follow this pattern. Observers do not see nearly as many small galaxies as the models predict, and the most massive galaxies appear to have formed quickly and efficiently, rather than through the gradual assembly of smaller pieces.

The models were clearly missing a critical ingredient. Something must be regulating the cooling of gas and the formation of stars in galaxies. The process has made small galaxies inefficient at forming stars and limited the size of massive galaxies. Theorists began considering a variety of additional physical processes that would provide this regulation. Known collectively as galactic feedback, these processes can counter, or reverse, the gravitational collapse of gas into galaxies and thus limit the number of stars that can form. They include supernovae explosions, stars' ultraviolet radiation and outflows, and the tremendous energy released during the growth of the supermassive black holes that lurk in the core of all massive galaxies [see "Black Hole Blowback," by Wallace Tucker, Harvey Tananbaum and Andrew Fabian; Scientific American, March 2007]. In the most massive galaxies, black holes are probably the most dominant feedback mechanism; in lower-mass systems, supernovae and stellar winds are more important.

What all these processes have in common is that they inject energy back into the surrounding medium. In this way, galaxies can choke off the inward flow of material, prevent gas that has already accumulated from forming stars or, in extreme cases, eject baryons back into intergalactic space. Simulations that take feedback into account do a much better job of reproducing the observed variety of galaxies. Not only does feedback play a critical role in tuning the evolution of galaxies, it can also resupply, reheat and enrich the WHIM. Through a continuous process of cooling and heating, baryons are cycled between intergalactic space and the stars and gas within galaxies. Galaxy growth is determined by a delicate balance of power that has tipped one way or another over cosmic history. Understanding this battle for baryons revamps our view of galaxy formation.

BLOBOLOGY

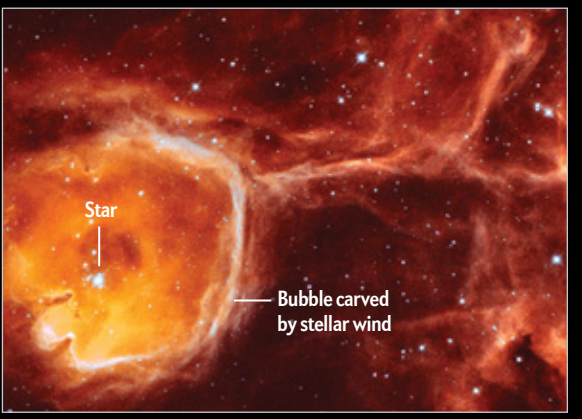
THE STUDY OF GAS COOLING AND FEEDBACK has been a major focus of astrophysics over the past decade. Without empirical data, we have no way of testing the models. Cold material flowing into galaxies in the early universe should give itself away by the diffuse glow that hydrogen emits as it cools. Feedback can be inferred from the bright infrared emission from intense star formation and x-ray or radio emission from the environs of a supermassive black hole. We need to catch both these processes in the act. Recently we may have done just that.

About 15 years ago Charles Steidel of the California Institute

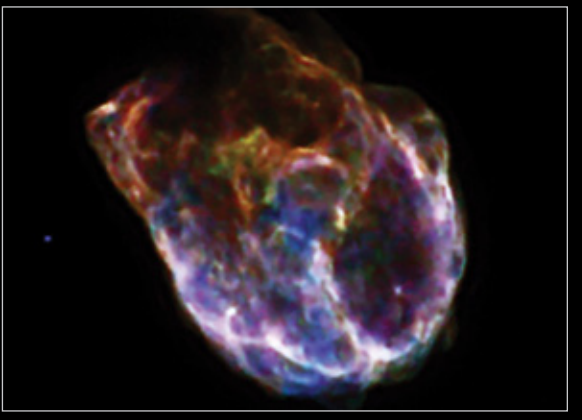
WHAT KEEPS MATTER OUT OF GALAXIES?

Spitting It Back Out

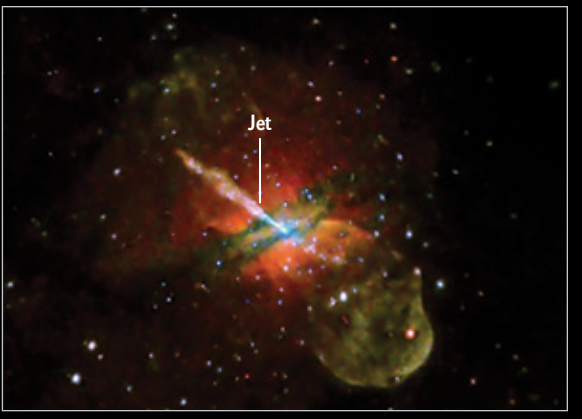
Why do galaxies account for for such a small fraction of cosmic matter? Not only does most of the matter end up in the WHIM, but galaxies also limit their own growth. The intergalactic gas that their gravity sucks in coalesces into stars and black holes. These objects feed energy back into the interstellar and intergalactic medium, countering the infall. Feedback processes include the winds blown out by stars, stellar explosions such as supernovae, and jets squirted out by black holes.



STAR



SUPERNOVA



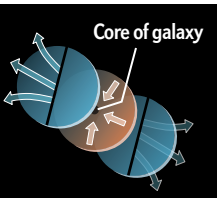
BLACK HOLE

COURTESY OF NASA/ESA/Y. NAZÉ *University of Liège* AND Y.-H. CHU *University of Illinois* (*star*); COURTESY OF NASA/CXC/NCSU/K. J. BORKOWSKI ET AL. (*supernova*); COURTESY OF NASA/CXC/CFA/R. KRAFT ET AL. (*black hole*)

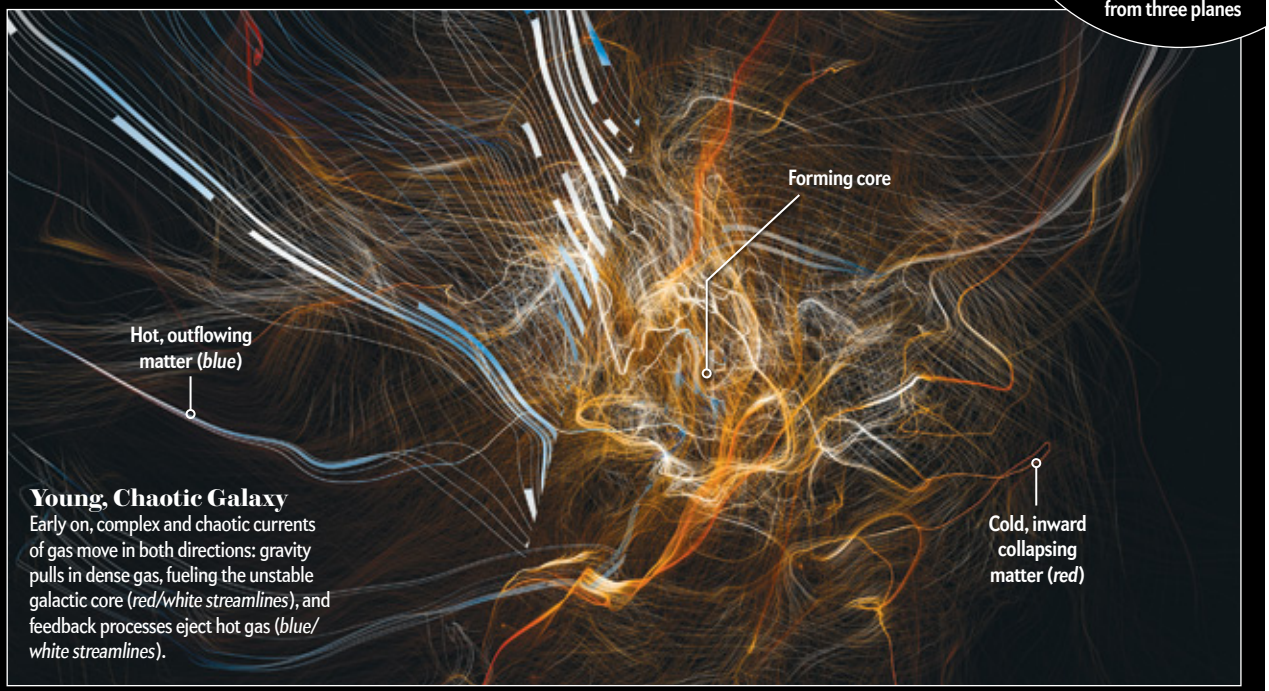
© 2014 Scientific American

The Turbulent Gyre

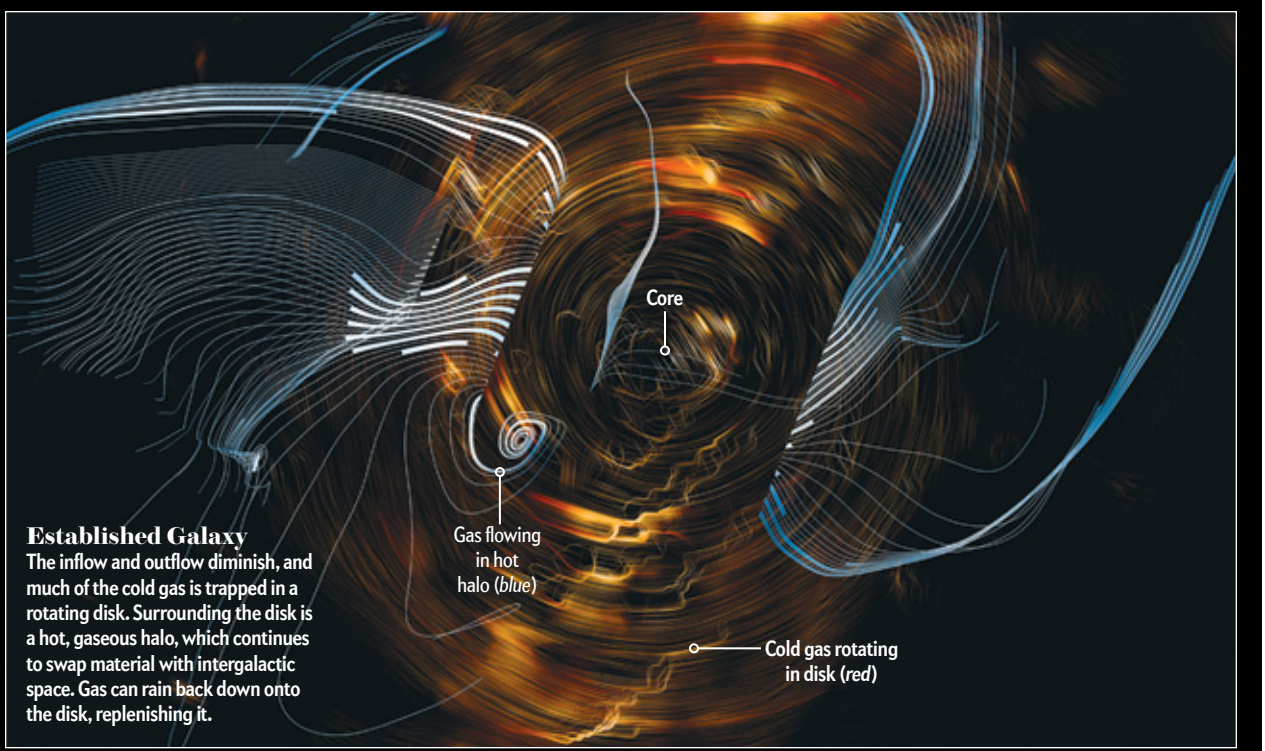
Most people think of galaxies as stately structures: giant balls or majestic pinwheels of stars floating in the emptiness of deep space. Astronomers are finding, however, that galaxies are fluid systems that actively exchange material with their surroundings. Normal matter cycles through galaxies, and at any given moment most of it resides in intergalactic space. This simulation shows a galaxy akin to our Milky Way as it would have appeared 10 billion years ago and today.



Visualizations below show streamlines emanating from three planes



Young, Chaotic Galaxy
Early on, complex and chaotic currents of gas move in both directions: gravity pulls in dense gas, fueling the unstable galactic core (*red/white streamlines*), and feedback processes eject hot gas (*blue/white streamlines*).



Established Galaxy
The inflow and outflow diminish, and much of the cold gas is trapped in a rotating disk. Surrounding the disk is a hot, gaseous halo, which continues to swap material with intergalactic space. Gas can rain back down onto the disk, replenishing it.

THE UNIVERSE

Graphics by James E. Geach

© 2014 Scientific American

of Technology and his collaborators discovered a new class of object that appears to tick the boxes for the observational signature of cooling: the Lyman-alpha blob. Never let it be said that astronomers are uptight with their nomenclature; "blob" really is the technical term. Lyman-alpha refers to one of the specific frequencies of light that hydrogen gas emits. These blobs appear to be glowing clouds up to 300,000 light-years across—far bigger than our Milky Way galaxy—making them among the largest luminous objects in the early universe. Astronomers have since discovered scores of them. The observed Lyman-alpha glow bears an uncanny resemblance to theoretical predictions of the radiative signature of cold gas flowing into young galaxies.

On the other hand, many other astrophysical processes could cause Lyman-alpha emission, too. For instance, ultraviolet light or a galactic-scale wind could pump energy into the blobs and cause them to glow. Using Chandra, my colleagues and I have shown that many blobs contain galaxies with actively growing black holes that shine brightly in x-rays. Often this activity is accompanied by intense star formation, revealed by its infrared emission from the obscuring layers of dust that blanket stellar birthing grounds. We have calculated that the energy released by these processes is more than enough to power the Lyman-alpha emission. So perhaps the blobs' glow is caused not by cooling, as many think, but by heating.

Rather than making things clearer, these blobs have muddied the waters somewhat. But that is what excites me about this field—it would not be science if we knew all the answers. We must now devise and conduct new observations to try to unveil what is really going on. Either way, though, blobs are precisely the type of object that could fill in some of the major gaps in our understanding about the origin of galaxies.

The observation that the intergalactic medium surrounding young, active galaxies gets swamped by radiation could help resolve another problem with galaxy-formation models. Very high resolution simulations of dark matter predict that galaxies such as the Milky Way should be accompanied by thousands of lower-mass dwarf galaxies buzzing around them like bees around a hive. Although the Milky Way does have a few dwarf companions, they are far fewer than the simulations predict.

One solution could be that the dwarf galaxies did form in the early universe, but their parent galaxy blasted them with radiation and winds. The barrage stripped away any baryons the dwarfs had managed to accumulate, leaving only barren clumps of dark matter that have skulked on the outskirts of the parent galaxy ever since. Larger galaxies reach a truce in the battle for baryons, but smaller ones lose the battle entirely.

WHAT IS A GALAXY?

PERHAPS THE MOST EXHILARATING experience a scientist can have is the feeling of a sea change in one's perspective on the world. For me, that came when I had to reevaluate what I thought of as "a galaxy." Traditionally we think of luminous galaxies as isolated and discrete island universes, as German philosopher Immanuel Kant put it. In some sense, that is clearly true. But the bright galactic islands of light are just the visible tips of a much wider, but still elusive, sea of baryonic matter. This material pervades the universe, distributed within and shaped by a vast, underlying dark architecture, continuously evolving through gravity.

All those baryons started off in the same state: a hot, pristine gas that rapidly formed the basic elements of hydrogen and helium, along with small amounts of deuterium and lithium. What we think of as galaxies formed from this raw material, pulled into dense concentration by gravity. But these structures are not fixed groups of baryons. Material moves among them as part of a vast cycle that has been in operation since the big bang. The competing influences of gravity and feedback cause gas to cool onto, and later get ejected from, galaxies. Recent computer simulations by Rob Crain and Benjamin D. Oppenheimer of Leiden University in the Netherlands and their collaborators suggest that up to half of the baryons currently locked into galaxies in the local universe have cycled through the intergalactic medium at least once and often many times. The baryons that make up your body have participated in this cycle for nearly 14 billion years; the matter within your fingernail could have formed in stars in other galaxies and then spent billions of years exiled in intergalactic space before coming to rest in our solar system. You are just an ephemeral phase, a brief host, to this rare substance we call "normal."

This concept of baryon cycling underpins the emerging view of galaxy evolution. The big picture you should have in mind is that galaxy evolution is just a small component of the large-scale evolution of the intergalactic medium. The baryonic universe is predominately gaseous, not galactic. The intergalactic medium is a battleground of forces, and amid this maelstrom, galaxies form. Galaxies are just one processing stage in a cycle that is continuously shifting baryons from one phase to the next, and at any one time, most of the baryons in the universe are not inside galaxies.

Sentimentally, perhaps, we hold galaxies in a special regard: the Milky Way is our cosmic habitat, a brilliant, vast, complex home within the darkness. From an anthropic viewpoint, we just happen to be lucky enough to exist at a time when the baryons that make up Earth and everything on it have taken on a cold, stable form. That will not always be the case. The death of the sun around five billion years from now will incinerate the inner planets, evaporate the outer ones and gradually disperse the resulting detritus of heavy elements back into the interstellar medium. Unless humans manage to cheat the cycle by developing the technical capability to escape the confines of the solar system, the ashes of every material thing on Earth are fated to be returned, enriched, to the cosmos. And so the cycle continues. SA

James E. Geach is an observational astronomer who studies evolution of galaxies, particularly their star-formation histories and the evolution of their cold gas. He traces his love of astronomy to the dark night skies of his native Cornwall, England. He is now a Royal Society University Research Fellow at the University of Hertfordshire, where he lives with his wife, Kristen, daughter, Sophie, and their beagle, Darwin.

MORE TO EXPLORE

The Search for the Missing Baryons at Low Redshift. Joel N. Bregman in *Annual Review of Astronomy and Astrophysics,* Vol. 45, No. 1, pages 221–259; September 2007. Preprint available at arxiv.org/abs/0706.1787

The Chandra Deep Protocluster Survey: Ly-Alpha Blobs Are Powered by Heating, Not Cooling. James E. Geach et al. in *Astrophysical Journal,* Vol. 700, No. 1, pages 1–9; July 2009. arxiv.org/abs/0904.0452

The Intergalactic Medium over the Last 10 Billion Years—I. Ly-Alpha Absorption and Physical Conditions. Romeel Davé et al. in *Monthly Notices of the Royal Astronomical Society,* Vol. 408, No. 4, pages 2051–2070; November 2010.

James E. Geach's Web site, with images and links: http://jgeach.github.io

scientificamerican.com/magazine/sa

© 2014 Scientific American

© 2014 Scientific American

COSMIC STRUCTURES

THE UNIVERSE'S INVISIBLE HAND

Dark energy does more than hurry along the expansion of the universe. It also has a stranglehold on the shape and spacing of galaxies

By Christopher J. Conselice

IN BRIEF

Best known as the putative agent responsible for the accelerating expansion of the universe, dark energy is an unidentified substance that exerts a kind of anti-gravity force on the universe as a whole.

But dark energy has other effects, such as shaping the distribution of matter on large scales. On smaller scales, it stunted the growth of galaxy clusters some six billion years ago and reduced the rate of galactic interactions.

The Milky Way might have had a lower star formation rate had dark energy been weaker or stronger—the heavy elements that constitute our planet might never have been synthesized.

Illustration by Jean-Francois Podevin

© 2014 Scientific American

WHAT TOOK US SO LONG? ONLY IN 1998 DID ASTRONOMERS DISCOVER we had been missing nearly three quarters of the contents of the universe, the so-called dark energy—an unknown form of energy that surrounds each of us, tugging at us ever so slightly, holding the fate of the cosmos in its grip, but to which we are almost totally blind. Some researchers, to be sure, had anticipated that such energy existed, but even they will tell you that its detection ranks among the most revolutionary discoveries in 20th-century cosmology. Not only does dark energy appear to make up the bulk of the universe, but its existence, if it stands the test of time, will probably require the development of new theories of physics.

Scientists are just starting the long process of figuring out what dark energy is and what its implications are. One realization has already sunk in: although dark energy betrayed its existence through its effect on the universe as a whole, it may also shape the evolution of the universe's inhabitants—stars, galaxies, galaxy clusters. Astronomers may have been staring at its handiwork for decades without realizing it.

Ironically, the very pervasiveness of dark energy is what made it so hard to recognize. Dark energy, unlike matter, does not clump in some places more than others; by its very nature, it is spread smoothly everywhere. Whatever the location—be it in your kitchen or in intergalactic space—it has the same density, about 10^{-26} kilogram per cubic meter, equivalent to a handful of hydrogen atoms. All the dark energy in our solar system amounts to the mass of a small asteroid, making it an utterly inconsequential player in the dance of the planets. Its effects stand out only when viewed over vast distances and spans of time.

DARK ENERGY MAY BE THE KEY LINK AMONG SEVERAL ASPECTS OF GALAXY FORMATION THAT USED TO APPEAR UNRELATED.

Since the days of American astronomer Edwin Hubble, observers have known that all but the nearest galaxies are moving away from us at a rapid rate. This rate is proportional to distance: the more distant a galaxy is, the faster its recession. Such a pattern implied that galaxies are not moving through space in the conventional sense but are being carried along as the fabric of space itself stretches [see "Misconceptions about the Big Bang," by Charles H. Lineweaver and Tamara M. Davis; Scientific American, March 2005]. For decades, astronomers struggled to answer the obvious follow-up question: How does the expansion rate change over time? They reasoned that it should be slowing down, as the inward gravitational attraction exerted by galaxies on one another should eventually counteract the outward expansion.

The first clear observational evidence for changes in the expansion rate involved distant supernovae, massive exploding stars that can be used as markers of cosmic expansion, just as watching driftwood lets you measure the speed of a river. These observations made clear that the expansion was slower in the past than today and is therefore accelerating. More specifically, it had been slowing down but at some point underwent a transition and began speeding up [see "Surveying Space-time with Supernovae," by Craig J. Hogan, Robert P. Kirshner and Nicholas B. Suntzeff; Scientific American, January 1999, and "From Slowdown to Speedup," by Adam G. Riess and Michael S. Turner; Scientific American, February 2004]. This striking result has since been cross-checked by independent studies of the cosmic microwave background radiation by, for example, the Wilkinson Microwave Anisotropy Probe (WMAP).

One possible conclusion is that different laws of gravity apply on supergalactic scales than on lesser ones, so that galaxies' gravity does not, in fact, resist expansion. But the more generally accepted hypothesis is that the laws of gravity are universal and that some form of energy, previously unknown to science, opposes and overwhelms galaxies' mutual attraction, pushing them apart ever faster. Although dark energy is inconsequential within our galaxy (let alone your kitchen), it adds up to the most powerful force in the cosmos.

© 2014 Scientific American

COSMIC SCULPTOR

AS ASTRONOMERS HAVE EXPLORED this new phenomenon, they have found that, in addition to determining the overall expansion rate of the universe, dark energy has long-term consequences for smaller scales. As you zoom in from the entire observable universe, the first thing you notice is that matter on cosmic scales is distributed in a cobweblike pattern—a filigree of filaments, several tens of millions of light-years long, interspersed with voids of similar size. Simulations show that both matter and dark energy are needed to explain the pattern.

That finding is not terribly surprising, though. The filaments and voids are not coherent bodies like, say, a planet. They have not detached from the overall cosmic expansion and established their own internal equilibrium of forces. Rather they are features shaped by the competition between cosmic expansion (and any phenomenon affecting it) and their own gravity. In our universe, neither player in this tug-of-war is overwhelmingly dominant. If dark energy were stronger, expansion would have won and matter would be spread out rather than concentrated in filaments. If dark energy were weaker, matter would be even more concentrated than it is [*see box on page 66*].

The situation gets more complicated as you continue to zoom in and reach the scale of galaxies and galaxy clusters. Galaxies, including our own Milky Way, do not expand with time. Their size is controlled by an equilibrium between gravity and the angular momentum of the stars, gas and other material that make them up; they grow only by accreting new material from intergalactic space or by merging with other galaxies. Cosmic expansion has an insignificant effect on them. Thus, it is not at all obvious that dark energy should have had any say whatsoever in how galaxies formed. The same is true of galaxy clusters, the largest coherent bodies in the universe—assemblages of thousands of galaxies embedded in a vast cloud of hot gas and bound together by gravity.

Yet it now appears that dark energy may be the key link among several aspects of galaxy and cluster formation that not long ago appeared unrelated. The reason is that the formation and evolution of these systems is partially driven by interactions and mergers between galaxies, which in turn may have been driven strongly by dark energy.

To understand the influence of dark energy on the formation of galaxies, first consider how astronomers think galaxies form. Current theories are based on the idea that matter comes in two basic kinds. First, there is ordinary matter, whose particles readily interact with one another and with electromagnetic radiation. Astronomers call this type of matter "baryonic" in reference to its main constituent, baryons, such as protons and neutrons. Second, there is dark matter (which is distinct from dark energy), which makes up 85 percent of all matter and whose salient property is that it comprises particles that do not react with radiation. Gravitationally, dark matter behaves just like ordinary matter.

According to models, dark matter began to clump immediately after the big bang, forming blobs of various shapes that astronomers refer to as "halos." The baryons, in contrast, were initially kept from clumping by their interactions with one another and with radiation. They remained in a hot, gaseous phase. As the universe expanded, this gas cooled and the baryons were able to pack themselves together. The first stars and

DISCOVERIES

Evidence for Dark Energy

Supernova Explosions
In an expanding universe, galaxies move apart at a speed that depends on the distance between them. Supernovae offer a way to measure this effect: their spectral redshift reveals the speed of their host galaxies, and their brightness reveals distance. It turns out that galaxies billions of years ago were moving slower than a simple extrapolation from the current rate of expansion would imply. The expansion rate must have increased over that time—the hallmark of dark energy.

Cosmic Microwave Background Radiation
Images of the background radiation contain spots whose apparent size reflects the overall geometry of space and therefore the density of the universe. This quantity exceeds the amount of matter (both ordinary and exotic), so a missing component such as dark energy must make up the difference. In addition, the background radiation has been slightly reworked by the gravitational fields of cosmic structures. The amount of reworking depends on how the expansion rate has changed over time and matches what dark energy would do.

Galaxy Configuration
Galaxies are not sprinkled randomly through the heavens. Instead they cluster in patterns with scalar features similar to those in the cosmic microwave background. Studying the scale of these patterns allows astronomers to measure various properties of dark energy.

Gravitational Lensing
A lump of mass can serve as a lens; its gravity bends light. Such a lens can produce multiple images, like a fun-house mirror, if the light source is directly behind it—an alignment that becomes more probable the bigger the universe is, which in turn depends on the amount of dark energy. A weaker lens can still bend light by a small angle that depends on its mass. Studies of this process have revealed how clumps of matter have grown over time and found the imprint of dark energy.

Galaxy Clusters
X-ray observations trace the evolution of galaxy clusters. Dark energy is required to explain their formation.

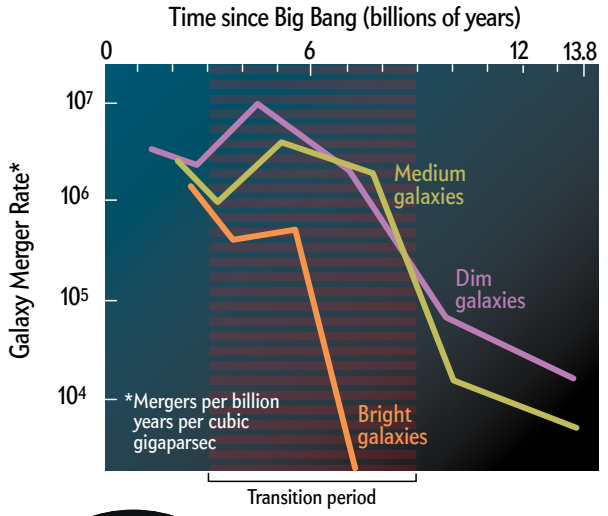
SUPERNOVA 1994D
in galaxy NGC 4526

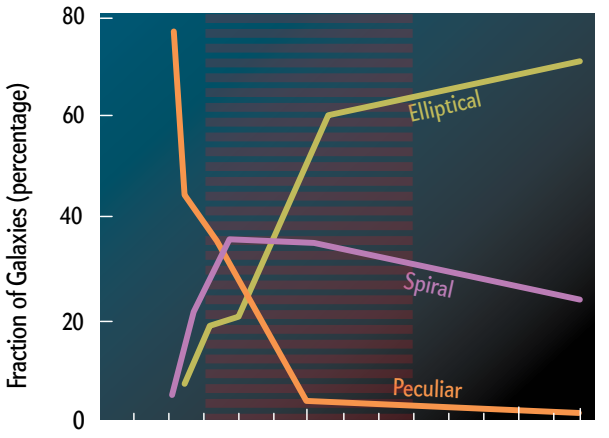
HIGH-Z SUPERNOVA SEARCH TEAM, HST, NASA

© 2014 Scientific American

Dark Energy Takes Charge

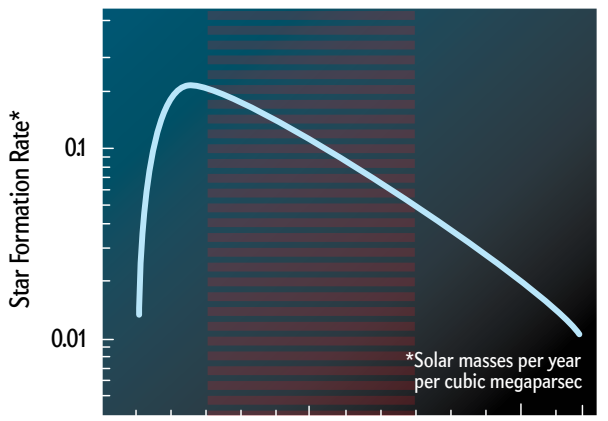
Despite getting a quick start after the big bang, the construction of the universe soon petered out. Initially galaxies merged together, changed shape and formed stars at a brisk pace, but this activity began to wane during the period when dark energy became comparable in strength to matter (*brown area in graphs*). Coincidence?

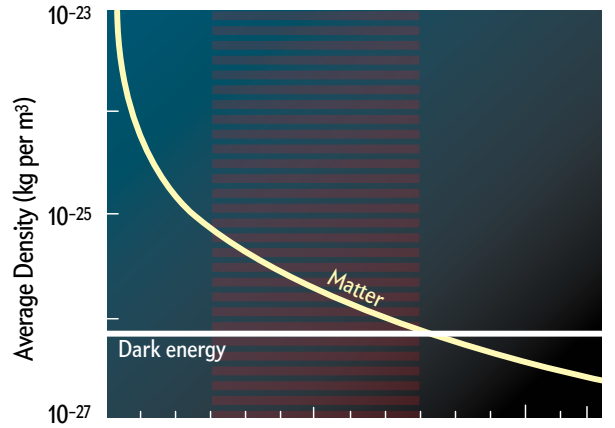




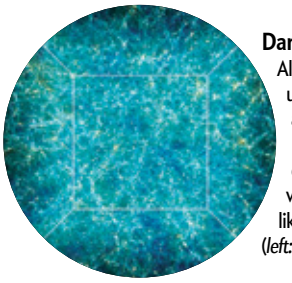
Galaxies Stopped Merging
The brightest galaxies stopped colliding and amalgamating at a cosmic age of about six billion years. Less luminous galaxies can still merge but have become much less likely to do so. (*left:* Merging galaxies NGC 4676)

Galaxies Settled into Regular Shapes
Early on, most galaxies looked peculiar—a sign they were merging with one another. As mergers became less frequent, spiral and elliptical shapes become prevalent. (*left:* Hickson Compact Galaxy Group 87)





Star Formation Waned
The early universe was a cauldron of star formation, but the formation rate soon peaked and began to drop. It is now lower than it has ever been. (*left:* New Stars in Trapezium Cluster)



Dark Energy Became a Player
All these trends relate to one fact: as the universe expanded and matter's density approached that of dark energy, the rate of expansion began to switch from decelerating to accelerating. Galaxies were pulled apart faster and became less likely to interact and undergo star formation. (*left:* Simulation of Matter Distribution)

LUCY READING-IKKANDA (*graphs*); ANDREW HOPKINS *University of Sydney (source of star formation rate graph)*; NASA/ACS SCIENCE & ENGINEERING TEAM *(merging galaxies)*; GMOS-S COMMISSIONING TEAM, GEMINI OBSERVATORY *(Hickson group)*; NASA/K. L. LUHMAN *Harvard-Smithsonian Center for Astrophysics* AND G. SCHNEIDER, E. YOUNG, G. RIEKE, A. COTERA, H. CHEN, M. RIEKE AND R. THOMPSON *Steward Observatory, University of Arizona (Trapezium cluster)*; IMAGE GENERATION BY VISUALIZATION INSTITUTE *University of Stuttgart*, DATA FROM VIRGO SUPERCOMPUTING CONSORTIUM *Max Planck Society, Garching, and Institute For Computational Cosmology (simulation of matter distribution)*

© 2014 Scientific American

galaxies coalesced out of this cooled gas a few hundred million years after the big bang. They did not materialize in random locations but in the centers of the dark matter halos that had already taken shape.

Since the 1980s a number of theorists have done detailed computer simulations of this process, including groups led by Simon D. M. White of the Max Planck Institute for Astrophysics in Garching, Germany, and Carlos S. Frenk of Durham University in England. They have shown that most of the first structures were small, low-mass dark matter halos. Because the early universe was so dense, these low-mass halos (and the galaxies they contained) merged with one another to form larger-mass systems. In this way, galaxy construction was a bottom-up process, like building a dollhouse out of Lego bricks. (The alternative would have been a top-down process, in which you start with the dollhouse and smash it to make bricks.) My colleagues and I have sought to test these models by looking at distant galaxies and how they have merged over cosmic time.

GALAXY FORMATION PETERS OUT

DETAILED STUDIES INDICATE that a galaxy gets bent out of shape when it merges with another galaxy. The earliest galaxies we can see existed when the universe was about a billion years old, and many of these indeed appear to be merging. As time went on, though, the fusion of massive galaxies became less common. Between two billion and six billion years after the big bang—that is, over the first half of cosmic history—the fraction of massive galaxies undergoing a merger dropped from half to just under 1 percent. Since then, the distribution of galaxy shapes has been frozen, an indication that smashups and mergers have become relatively uncommon.

THE UNIVERSE HAS RUN OUT OF STEAM SINCE IT WAS HALF ITS CURRENT AGE. MERGERS HAVE CEASED, AND BLACK HOLES ARE QUIESCENT.

In fact, fully 98 percent of massive galaxies in today's universe are either elliptical or spiral—shapes and structures that would be disrupted by a merger. These galaxies are stable and comprise mostly old stars, which tells us that they must have formed early and have remained in a regular morphological form for quite some time. A few galaxies are merging in the present day, but they are typically of low mass.

The virtual cessation of mergers is not the only way the universe has run out of steam since it was half its current age. Star formation, too, has been waning. Most of the stars that exist today were born in the first half of cosmic history, as first convincingly shown by several teams in the 1990s, including ones led by Simon J. Lilly, then at the University of Toronto, Piero Madau, then at the Space Telescope Science Institute, and Charles C. Steidel of the California Institute of Technology. More recently, researchers have learned how this trend occurred. It turns out that star formation in massive galaxies shut down early. Since the universe was half its current age, only lightweight systems have continued to create stars at a significant rate. This shift in star formation is called galaxy downsizing [see "The Midlife Crisis of the Cosmos," by Amy J. Barger; SCIENTIFIC AMERICAN, January 2005]. It seems paradoxical and is one of the biggest mysteries in extragalactic astronomy. Theory predicts that small galaxies take shape first, and as they merge, massive ones arise. Yet the history of star formation shows the reverse: massive galaxies are initially the main stellar birthing grounds, then smaller ones take over.

Another oddity is that the buildup of supermassive black holes, found at the centers of galaxies, seems to have slowed down considerably. Such holes power quasars and other types of active galaxies, which are rare in the modern universe; the black holes in our galaxy and others are quiescent. Are any of these trends in galaxy evolution related? Is it really possible that dark energy is the root cause?

THE STEADY GRIP OF DARK ENERGY

SOME ASTRONOMERS have proposed that internal processes in galaxies, such as energy released by black holes, supernovae and stellar winds, turned off galaxy and star formation. But dark energy has emerged as possibly a more fundamental culprit, the one that can link everything together. The central piece of evidence is the rough coincidence in timing between the end of most galaxy, black hole and cluster formation and the onset of the domination of dark energy. Both happened when the universe was about half its present age.

The idea is that up to that point in cosmic history, the density of matter was so high that gravitational forces among galaxies dominated over the effects of dark energy. Galaxies rubbed shoulders, interacted with one another and frequently merged. New stars formed as gas clouds within galaxies collided, and black holes grew when gas was driven toward the centers of these systems. As time progressed and space expanded, matter thinned out and its gravity weakened, whereas the strength of dark energy remained constant or increased. The inexorable shift in the balance between the two eventually caused the expansion rate to switch from deceleration to acceleration. The structures in which galaxies reside were then pulled apart, with a gradual decrease in the galaxy merger rate as a result. Likewise, intergalactic gas was less able to fall into galaxies. Deprived of fuel, black holes became more quiescent.

This sequence could perhaps account for the downsizing of the galaxy population. The most massive dark matter halos, as well as their embedded galaxies, are also the most clustered; they reside in close proximity to other massive halos. Thus, they are likely to knock into their neighbors earlier than are lower-mass systems. When they do, they experience a burst of star formation. The newly formed stars light up and then blow up, heating the gas and preventing it from collapsing into new stars. In this way, star formation chokes itself off: stars heat the gas from which they emerged, preventing new ones from forming. The black hole at the center of such a galaxy acts as another damper on star formation. A galaxy merger feeds gas into the black hole, causing it to fire out jets that heat up gas in the system and prevent it from cooling to form new stars.

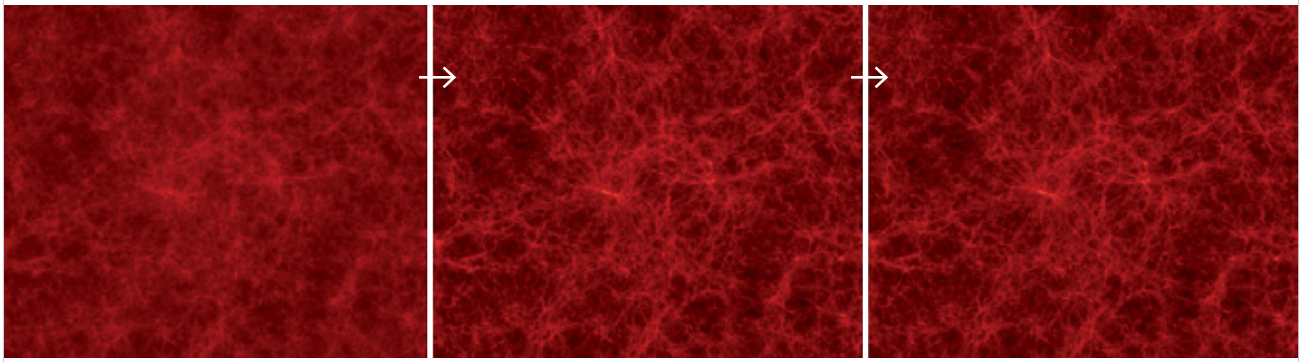
Apparently, once star formation in massive galaxies shuts down, it does not start up again—most likely because the gas in these systems becomes depleted or becomes so hot that it cannot cool down quickly enough. These massive galaxies can still merge with one another, but few new stars emerge for want of

© 2014 Scientific American

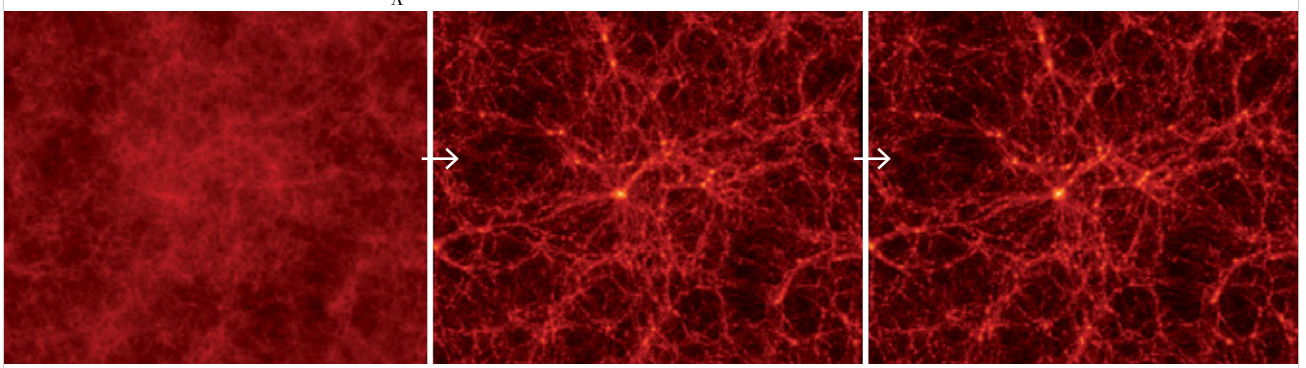
What-If Scenarios

If the universe had more dark energy in it, it would look radically different. Cosmic acceleration would have started sooner, pulled material apart faster and nipped the formation of large structures in the bud. The converse would happen if the universe had less dark energy. Each box below shows a region that is now one billion light-years across and contains 27 million particles, each representing a galaxy. These simulations assume that the dark energy density is constant in space and time. The quantity Ω_Λ is the governing cosmological parameter; it represents the density of dark energy today.

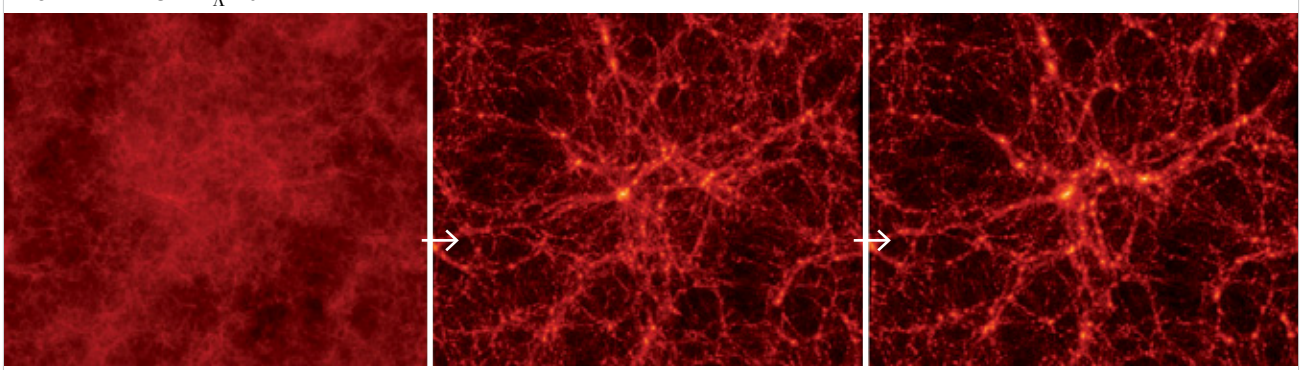
MORE DARK ENERGY $\Omega_\Lambda = 0.99$



OBSERVED AMOUNT OF DARK ENERGY $\Omega_\Lambda = 0.75$



NO DARK ENERGY $\Omega_\Lambda = 0$



EARLY UNIVERSE: When the universe is a sixth of its current size, matter is evenly distributed in all three scenarios. Dark energy has not yet exerted its influence.

TRANSITION PERIOD: When the universe is 75 percent of its current size, the effects of dark energy are stark. In the high dark energy scenario (*top*), the universe looks amorphous. In the other two scenarios, structure formation still continues, producing a cobweb pattern.

TODAY: In a universe with the observed amount of dark energy (*middle*), large-scale structure formation has ended, leaving the cobweb frozen in place. In a zero dark energy scenario (*bottom*), the cobweb continues to develop.

LIANG GAO, CARLOS FRENK AND ADRIAN JENKINS *Institute for Computational Cosmology, Durham University*

© 2014 Scientific American

cold gas. As the massive galaxies stagnate, smaller galaxies continue to merge and form stars. The result is that massive galaxies take shape before smaller ones, as is observed. Dark energy perhaps modulated this process by determining the degree of galaxy clustering and the rate of merging.

Dark energy would also explain the evolution of galaxy clusters. Ancient clusters, found when the universe was less than half its present age, were already as massive as today's clusters. That is, galaxy clusters have not grown by a significant amount in the past six billion to eight billion years. This lack of growth is an indication that the infall of galaxies into clusters has been curtailed since the universe was about half its current age—a direct sign that dark energy is influencing the way galaxies are interacting on large scales. Astronomers knew as early as the mid-1990s that galaxy clusters had not grown much in the past eight billion years, and they attributed this to a lower matter density than theoretical arguments had predicted. The discovery of dark energy helped to resolve the tension between observation and theory.

An example of how dark energy alters the history of galaxy clusters is the fate of the galaxies in our immediate vicinity, known as the Local Group. Several years ago astronomers thought that the Milky Way and Andromeda, its closest large neighbor, along with their retinue of satellites, would eventually fall into the nearby Virgo cluster. But it now appears that we shall escape that fate and never become part of a large cluster of galaxies. Dark energy will cause the distance between us and Virgo to expand faster than the Local Group can cross it.

By throttling cluster development, dark energy also controls the makeup of galaxies within clusters. The cluster environment facilitates the formation of a zoo of galaxies such as the so-called lenticulars, giant ellipticals and dwarf ellipticals. By regulating the ability of galaxies to join clusters, dark energy dictates the relative abundance of these galaxy types.

This is a good story, but is it true? Galaxy mergers, black hole activity and star formation all decline with time, and very likely they are related in some way. But astronomers have yet to follow the full sequence of events. Ongoing surveys with the Hubble Space Telescope, the Chandra X-ray Observatory and sensitive ground-based imaging and spectroscopy may deliver answers in coming years. One way to do this is to obtain a good census of distant active galaxies and to determine the time when those galaxies last underwent a merger. The analysis will require the development of new theoretical tools but should be within our grasp in the not too distant future.

STRIKING A BALANCE

AN ACCELERATING UNIVERSE dominated by dark energy is a natural way to produce all the observed changes in the galaxy population—namely the cessation of mergers and its many corollaries, such as loss of vigorous star formation and the end of galactic metamorphosis. If dark energy did not exist, galaxy mergers would probably have continued for longer than they did, and today the universe would contain many more massive galaxies with old stellar populations. Likewise, it would have fewer lower-mass systems, and spiral galaxies such as our Milky Way would be rare (given that spirals cannot survive the merger process). Large-scale structures of galaxies would have been more tightly bound, and more mergers of structures and accretion would have occurred.

Conversely, if dark energy were even stronger than it is, the universe would have had fewer mergers and thus fewer massive galaxies and galaxy clusters. Spiral and low-mass dwarf irregular galaxies would be more common because fewer galaxy mergers would have occurred throughout time, and galaxy clusters would be much less massive or perhaps not exist at all. It is also likely that fewer stars would have formed, and a higher fraction of our universe's baryonic mass would still be in a gaseous state.

SPACE IS EMPTYING OUT, LEAVING OUR MILKY WAY GALAXY AND ITS NEIGHBORS AN INCREASINGLY ISOLATED ISLAND.

Although these processes may seem distant from us here on Earth, the way galaxies form has an influence on our own existence. Stars are needed to produce elements heavier than lithium, which are used to build terrestrial planets and life. If lower star formation rates meant that these elements did not form in great abundance, the universe would not have many planets, and life itself might never have arisen. In this way, dark energy could have had a profound effect on many different and seemingly unrelated aspects of the universe and perhaps even on the detailed history of our own planet.

Dark energy is by no means finished with its work. It may appear to benefit life: the acceleration will prevent the eventual collapse that was a worry of astronomers not so long ago. But dark energy brings other risks. At the very least, it pulls apart distant galaxies, making them recede so fast that we lose sight of them for good. Space is emptying out, leaving our galaxy and its immediate neighbors an increasingly isolated island. Galaxy clusters, galaxies and even stars drifting through intergalactic space will eventually have a limited sphere of gravitational influence not much larger than their own individual sizes.

Worse, dark energy might be evolving. Some models predict that if dark energy becomes ever more dominant over time, it will rip apart gravitationally bound objects, such as galaxy clusters and galaxies. Ultimately, planet Earth will be stripped from the sun and shredded, along with all objects on it. Even atoms will be destroyed. Dark energy, once cast in the shadows of matter, will have exacted its final revenge. SA

Christopher J. Conselice is an astronomer and professor at the University of Nottingham in England. He specializes in the formation of galaxies and leads several observational programs in infrared and visible light with telescopes both on the ground and in space.

MORE TO EXPLORE

A Direct Measurement of Major Galaxy Mergers at z ≲ 3. Christopher J. Conselice, Matthew A. Bershady, Mark Dickinson and Casey Papovich in *Astronomical Journal*, Vol. 126, No. 3, pages 1183–1207; September 2003. www.arxiv.org/abs/astro-ph/0306106

Dark Energy. Robert R. Caldwell in *Physics World*, Vol. 17, No. 5, pages 37–42; May 2004. http://physicsworld.com/cws/article/print/2004/may/30/dark-energy

The Extravagant Universe: Exploding Stars, Dark Energy, and the Accelerating Cosmos. Robert P. Kirshner. Princeton University Press, 2004.

The Infinite Cosmos: Questions from the Frontiers of Cosmology. Joseph Silk. Oxford University Press, 2006.

scientificamerican.com/magazine/sa

© 2014 Scientific American

THEORETICAL PHYSICS

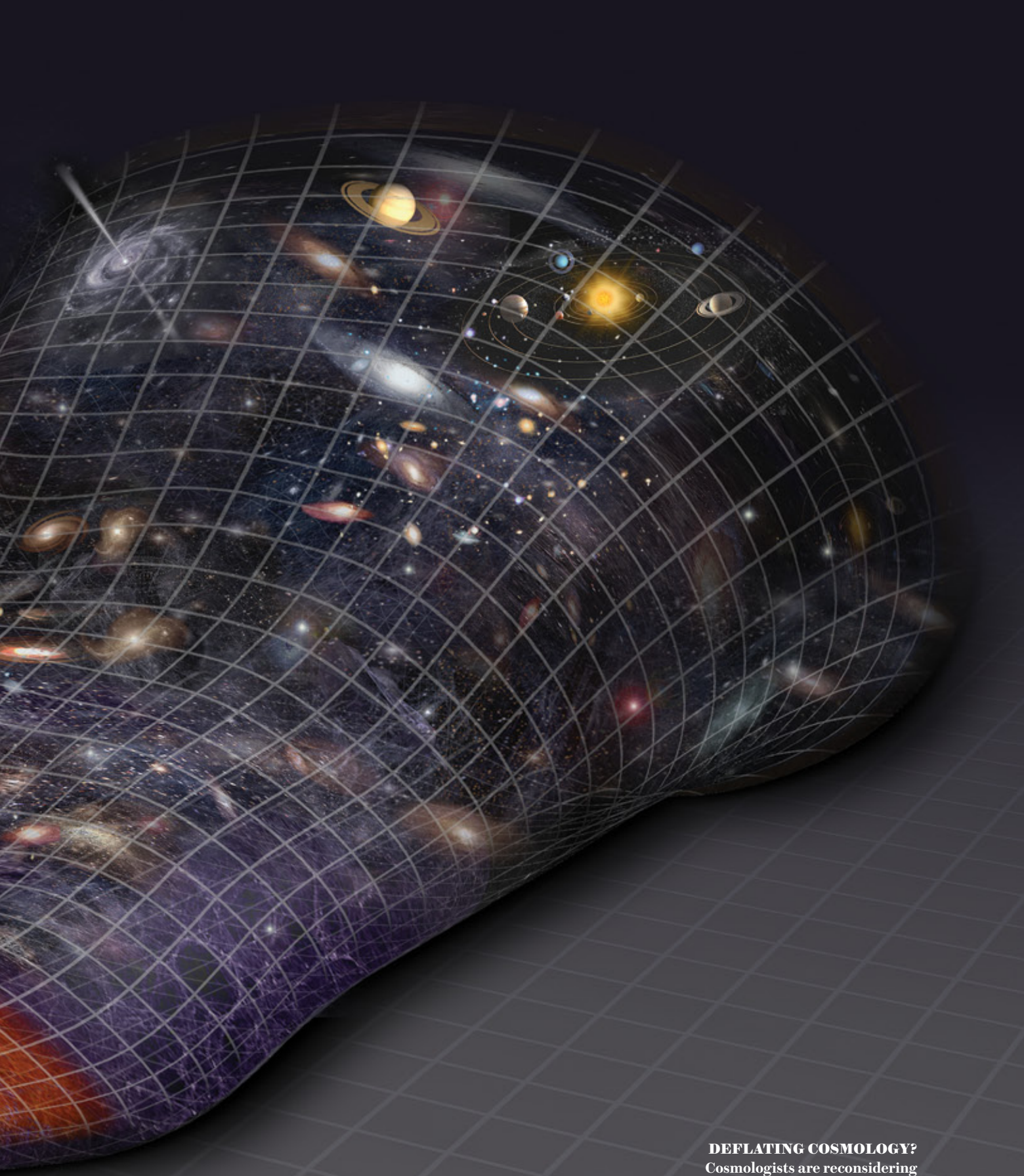
THE INFLATION DEBATE

Is the theory at the heart of modern cosmology deeply flawed?

By Paul J. Steinhardt

Illustrations by Malcolm Godwin

© 2014 Scientific American



DEFLATING COSMOLOGY? Cosmologists are reconsidering whether the universe really went through an intense growth spurt (*yellowish region*) shortly after the big bang.

© 2014 Scientific American

BACK IN 1981, ALAN H. GUTH, THEN A STRUGGLING PHYSICS POSTDOC AT the Stanford Linear Accelerator Center, gave a series of seminars in which he introduced "inflation" into the lexicon of cosmology. The term refers to a brief burst of hyperaccelerated expansion that, he argued, may have occurred during the first instants after the big bang. One of these seminars took place at Harvard University, where I myself was a postdoc. I was captivated by the idea, and I have been thinking about it almost every day since. Many of my colleagues working in astrophysics, gravitational physics and particle physics have been similarly engrossed. To this day the development and testing of the inflationary theory of the universe is one of the most active and successful areas of scientific investigation.

Its raison d'être is to fill a gap in the original big bang theory. The basic idea of the big bang is that the universe has been slowly expanding and cooling ever since it began some 13.8 billion years ago. This process explains many of the detailed features of the universe seen today, but with a catch: the universe had to start off with certain properties. For instance, it had to be extremely uniform, with only extremely tiny variations in the distribution of matter and energy. Also, the universe had to be geometrically flat, meaning that curves and warps in the fabric of space did not bend the paths of light rays and moving objects.

But why should the primordial universe have been so uniform and flat? A priori, these starting conditions seemed unlikely. That is where Guth's idea came in. He argued that even if the universe had started off in total disarray—with a highly nonuniform distribution of energy and a gnarled shape—a spectacular growth spurt would have spread out energy until it was evenly dispersed and straightened out any curves and warps in space. When this period of inflation ended, the universe would have continued to expand at the more mellow pace of the original big bang theory but now with just the right conditions for stars and galaxies to evolve to the state where we see them today.

The idea is so compelling that cosmologists, including me, routinely describe it as an established fact. Yet something peculiar has happened to inflationary theory in the years since Guth introduced it. As the case for inflation has grown stronger, so has the case against. The two cases are not equally well known: the evidence favoring inflation is familiar to a broad range of physicists, astrophysicists and science aficionados. Yet few seem to follow the case against inflation except for a small group of us who have been quietly striving to address the challenges. Most astrophysicists have gone about their business testing the predictions of textbook inflationary theory without worrying about these deeper issues, hoping they would eventually be resolved. Unfortunately, the problems have resisted our best efforts to date.

As someone who has contributed both to inflationary theory [see "The Inflationary Universe," by Alan H. Guth and Paul J. Steinhardt; SCIENTIFIC AMERICAN, May 1984] and to competing theories, I feel torn, and I sense that many of my colleagues are not sure what to make of the case against, either. To dramatize our strange predicament, I will place inflationary cosmology on trial, presenting the two extreme points of view. First, I will act as fervent advocate "for," presenting the strongest advantages of the theory, and then, with equal fervor, as advocate "against," presenting the most serious unresolved problems.

THE CASE FOR INFLATION

INFLATION is so well known that the case for it can be brief. A few more details are necessary to appreciate its advantages fully. Inflation relies on a special ingredient known as inflationary energy, which, combined with gravity, can drive the universe to expand by an astonishing amount over a brief instant. The inflationary energy must be hugely dense, and its density must

IN BRIEF

Cosmic inflation is so widely accepted that it is often taken as established fact. The idea is that the geometry and uniformity of the cosmos were established during an intense early growth spurt.

But some of the theory's creators, including the author, are having second thoughts. As the original theory has developed, cracks have appeared in its logical foundations.

Highly improbable conditions are required to start inflation. Worse, inflation goes on eternally, producing infinitely many outcomes, so the theory makes no firm observational predictions.

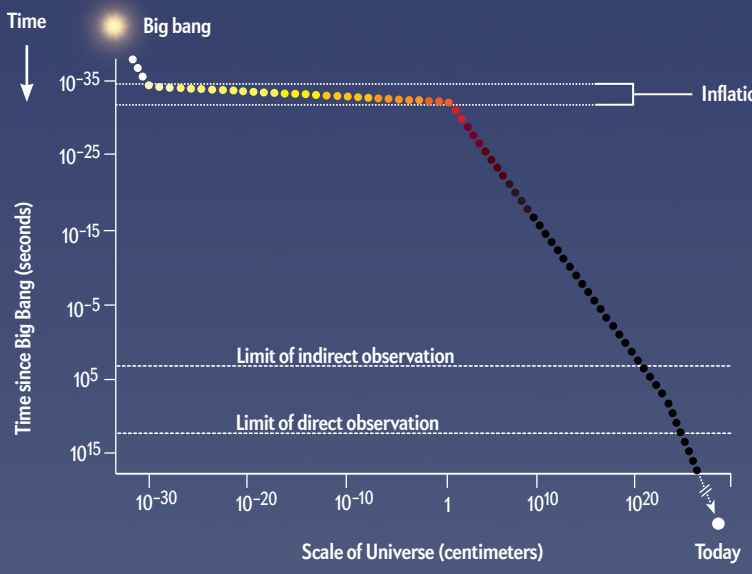
Scientists debate among (and within) themselves whether these troubles are teething pains or signs of a deeper rot. Various proposals are circulating for ways to fix inflation or replace it.

© 2014 Scientific American

The Ultimate Growth Spurt

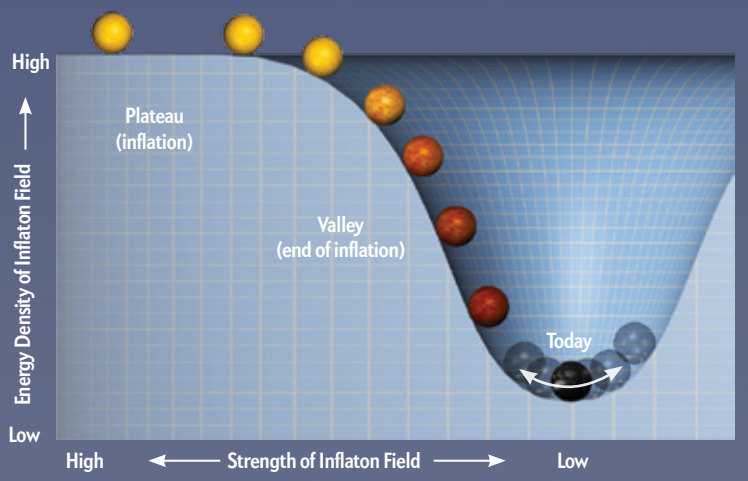
Astronomers observe that the universe is expanding and has been doing so for 13.8 billion years. But what happened at the very earliest times, too early to see directly? The leading idea is known as cosmic inflation. It supposes that the embryonic universe abruptly ballooned in size. Such a growth spurt would have ironed out any curves and warps in space, thus explaining the geometry of the universe today, and left behind slight nonuniformities that seeded galaxies.

WHAT INFLATION DID

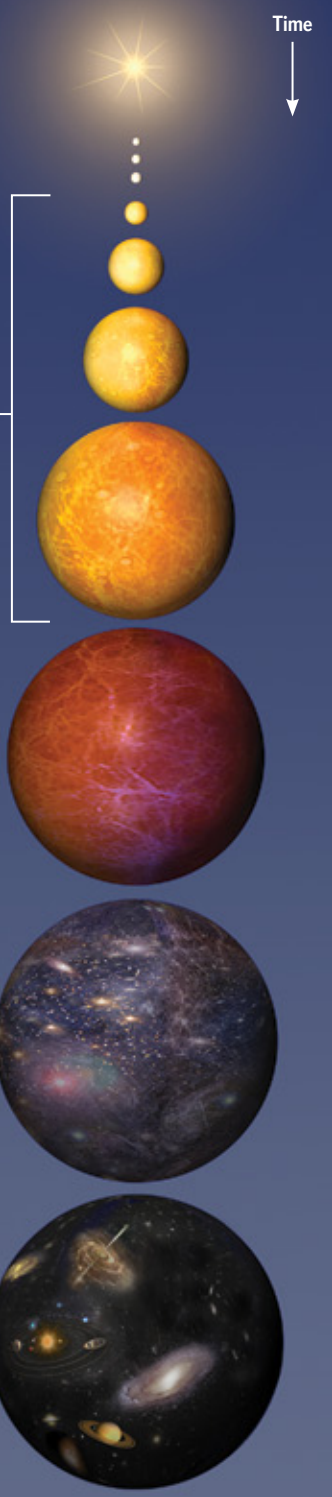


The amount of growth was impressive even by astronomers' standards. Within 10^{-30} second, the universe enlarged by a factor of at least 10^{25} in every direction. It expanded at an accelerated rate, pulling regions of space apart faster than the speed of light.

WHAT CAUSED INFLATION



A relative of the magnetic field, the "inflaton" generated a repulsive gravitational force that drove space to swell rapidly momentarily. For that to occur, the field's energy density had to vary with strength such that it had a high-energy plateau and a low-energy valley. The field evolved like a ball rolling downhill. On the plateau, it exerted the repulsive force. When it hit the valley, inflation ended.



The volume of space we observe today was a quadrillionth the size of an atom when inflation began. During inflation it grew to the size of a dime. In the billions of years since then, space has continued to expand but at a mellower pace, allowing structures such as galaxies to form. (This figure is conceptual and not to scale.)

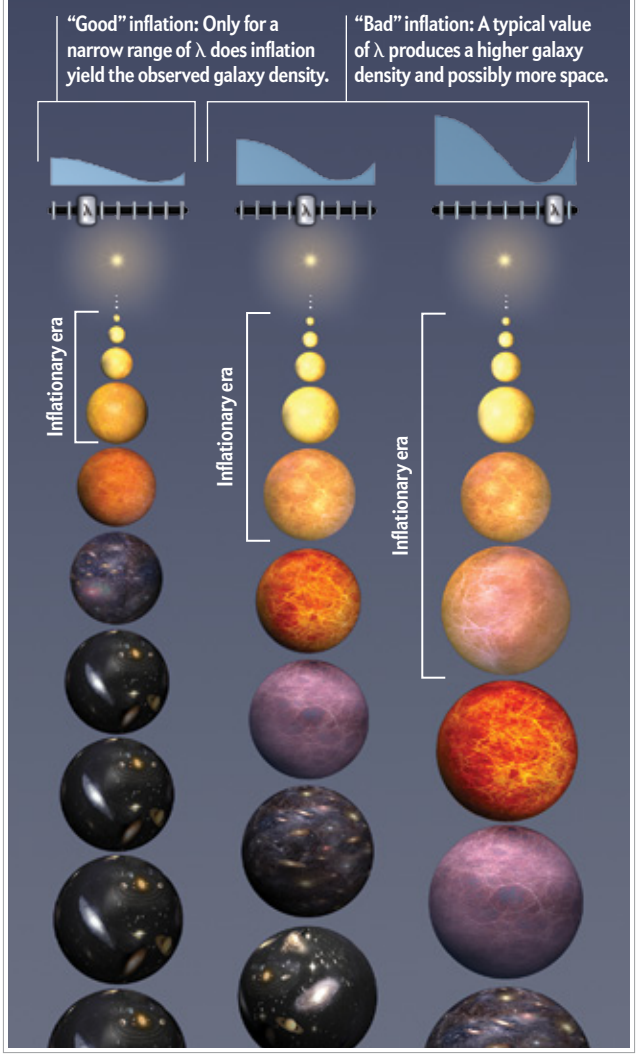
Graphic by Jen Christiansen (top left)

© 2014 Scientific American

PROBLEM #1: "BAD" INFLATION

Unlikely to Be Good

Inflation was supposed to create a huge volume of space matching the observed large-scale features of our universe naturally. But unless the inflaton energy curve had a specific shape (obtained by finely tuning one or more parameters, abbreviated λ here), the outcome would be "bad"—a volume with too high a density and the wrong distribution of galaxies. Given the range of possible λ values, bad inflation seems more likely.


"Good" inflation: Only for a narrow range of λ does inflation yield the observed galaxy density.
"Bad" inflation: A typical value of λ produces a higher galaxy density and possibly more space.
λ
Inflationary era
Inflationary era
Inflationary era

remain nearly constant during the inflationary epoch. Its most unusual property of all is that its gravity must repel rather than attract. The repulsion is what causes space to swell so rapidly.

What gave Guth's idea its appeal was that theorists had already identified many possible sources of such energy. The leading example is a hypothesized relative of the magnetic field known as a scalar field, which, in the case of inflation, is known as the "inflaton" field. The Higgs particle now being studied at CERN's Large Hadron Collider derives from another scalar field.

Like all fields, the inflaton has a certain strength at every point in space, which determines the force it exerts. During the inflationary phase, its strength is nearly constant everywhere. Depending on how strong a field is, it has a certain amount of energy in it—what physicists call potential energy. The relation between the strength and the energy can be shown by a curve on a graph. For the inflaton, cosmologists believe the curve looks like the cross section through a valley and a gently sloped plateau [*see box on preceding page*]. If the field begins with a strength corresponding to some point on the plateau, it will gradually lose both strength and energy, as if sliding down the slope. In fact, the equations are similar to those of a ball rolling down a hill of the same shape as the potential energy curve.

The inflaton's potential energy can cause the universe to expand at an accelerated rate. This can smooth and flatten the universe if the field stays on the plateau long enough (about 10^{-30} second) to stretch the universe by a factor of 10^{25} or more along each direction. Inflation ends when the field leaves the plateau and rushes downhill to the energy valley below. At this point, the potential energy converts into more familiar forms of energy—namely dark matter, hot ordinary matter and radiation—and the universe enters a period of modest, decelerating expansion.

Inflation smoothes the universe just as stretching a rubber sheet smoothes its wrinkles, but it does not do so perfectly. Small irregularities remain because of quantum effects. The laws of quantum physics dictate that a field such as the inflaton not have exactly the same strength everywhere in space but that it undergo random fluctuations. These fluctuations cause inflation to end at slightly different times in different regions of space, heating them to slightly different temperatures. These spatial variations are the seeds that will eventually grow into stars and galaxies. A prediction of inflationary theory is that the variations are nearly scale-invariant. That is, they do not depend on the size of the region; they occur with equal magnitude on all scales.

The case for inflation can be summarized by three dictums. First, inflation is inevitable. Recent developments in theoretical physics have only strengthened the hypothesis that the early universe contained fields to drive inflation. Hundreds appear in unified theories, such as string theory. In the chaotic primeval universe, there was sure to be some patch of space where one of these fields met the conditions for inflation.

Second, inflation explains why the universe is so uniform and flat today. No one knows how uniform or flat the universe was when it emerged from the big bang, but with inflation there is no need to know because the period of accelerated expansion stretched it into the right shape.

Third, and probably the most compelling, inflationary theory is powerfully predictive. For example, numerous observations of the cosmic microwave background radiation and the distribution of galaxies have confirmed that the spatial variations in energy in the early universe were nearly scale-invariant.

THE CASE AGAINST INFLATION

THE FIRST SIGNS that a theory is failing are usually small discrepancies between observations and predictions. That is not the situation here: the data are in exquisite accord with the inflationary predictions set down in the early 1980s. Instead the case against inflation challenges the logical foundations of the theory. Does the theory really work as advertised? Are the predictions made in the early 1980s still the predictions of the inflationary model as we understand it today? There is an

© 2014 Scientific American

argument to be made that the answer to both questions is no.

The first dictum holds that inflation is inevitable. But if it is, there is an awkward corollary: bad inflation is more likely than good inflation. "Bad" inflation means a period of accelerated expansion whose outcome conflicts with what we see. For example, the temperature variations might be too large. The difference between good and bad hinges on the shape of the potential energy curve, which is controlled by a numerical parameter that could, in principle, take on any value. Only a very narrow range of values could produce the observed temperature variation. In a typical case, the value must be near 10^{-15}—that is, zero to 15 decimal places. A less fine-tuned choice, such as zero to only 12 or 10 decimal places, would produce bad inflation, with a large temperature variation that is inconsistent with observations.

We could ignore bad inflation if it were incompatible with life. In that case, even if such large temperature variations could arise, we could never observe them. This reasoning is known as the anthropic principle, yet it does not apply here. Larger temperature variations would result in more stars and galaxies—the universe would be more habitable than it is now.

Not only is bad inflation more likely than good inflation, but no inflation is more likely than either. University of Oxford physicist Roger Penrose first made this point in the 1980s. He applied thermodynamic principles, similar to those used to describe atoms and molecules in a gas, to count the possible starting configurations of the inflaton and gravitational fields. Some of these configurations lead to inflation and thence to a nearly uniform, flat universe. Other configurations lead to a uniform, flat universe directly—without inflation. Both sets of configurations are rare, so obtaining a flat universe is unlikely overall. Penrose's shocking conclusion, though, was that obtaining a flat universe without inflation is much more likely than with inflation—by a factor of 10 to the googol (10^{100}) power!

THE PERILS OF AN ETERNAL INFLATION

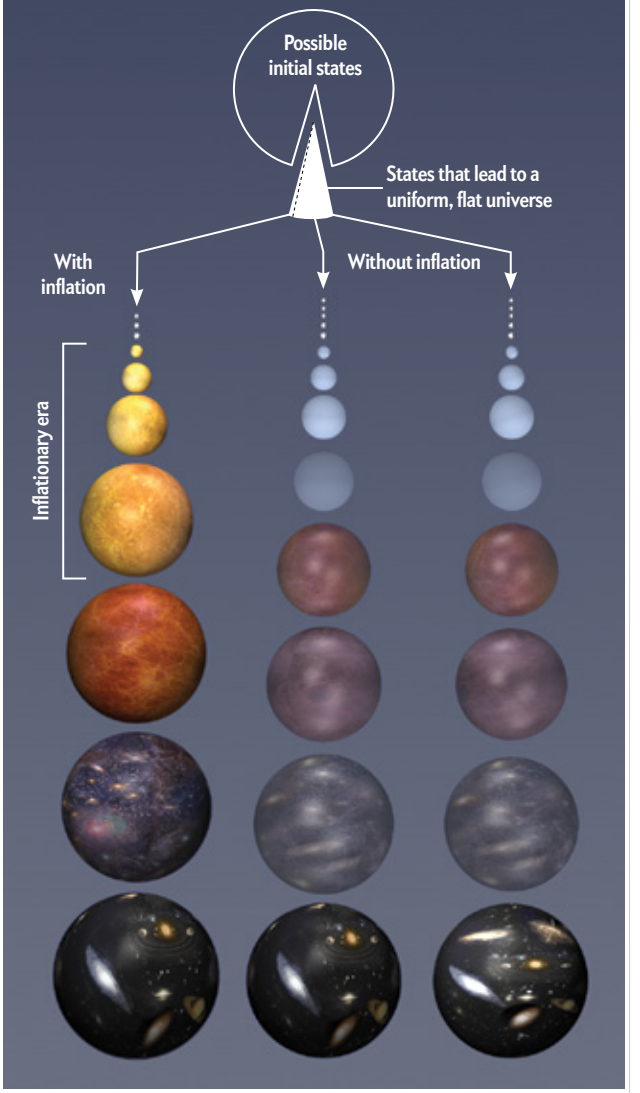
ANOTHER APPROACH extrapolates the history of the universe from its current conditions backward in time using the established physical laws. The extrapolation is not unique: given the average flat and smooth conditions today, many different sequences of events could have come before. In 2008 Gary W. Gibbons of the University of Cambridge and Neil G. Turok of the Perimeter Institute for Theoretical Physics in Ontario showed that an overwhelming number of extrapolations have insignificant amounts of inflation. This conclusion is consistent with Penrose's. Both seem counterintuitive because a flat and smooth universe is unlikely, and inflation is a powerful mechanism for obtaining the needed smoothing and flattening. Yet this advantage appears to be completely offset by the fact that the conditions for starting inflation are so improbable. When all factors are taken into account, the universe is more likely to have achieved its current conditions without inflation than with it.

Many researchers find these arguments unconvincing compared with a more compelling one favoring inflation: namely the agreement between the predictions formulated in the early 1980s and the cosmological observations available today. Corroborating evidence trumps any theoretical argument. But the strange twist to this story is that the predictions of the early 1980s were based on a naive understanding of how inflation actually works—a picture that has turned out to be dead wrong.

PROBLEM #2: STARTING CONDITIONS

It Had to Be Just So

Inflation was supposed to occur no matter what the initial conditions of the universe were. Further analysis suggests otherwise. Of all the ways the universe could have begun, only a tiny fraction would lead to the uniform, flat state observed today. An overwhelming fraction of these would reach this state without significant inflation; only an infinitesimal fraction would do so by going through a long period of inflation.



The change in view began with the realization that inflation is eternal: once begun, it never ends [see "The Self-Reproducing Inflationary Universe," by Andrei Linde; SCIENTIFIC AMERICAN, November 1994]. The self-perpetuating nature of inflation is the direct result of quantum physics combined with accelerated expansion. Recall that quantum fluctuations can delay when inflation ends. Where these fluctuations are small, so are their effects. Yet the fluctuations are random. In some regions of space, they will be large, leading to substantial delays.

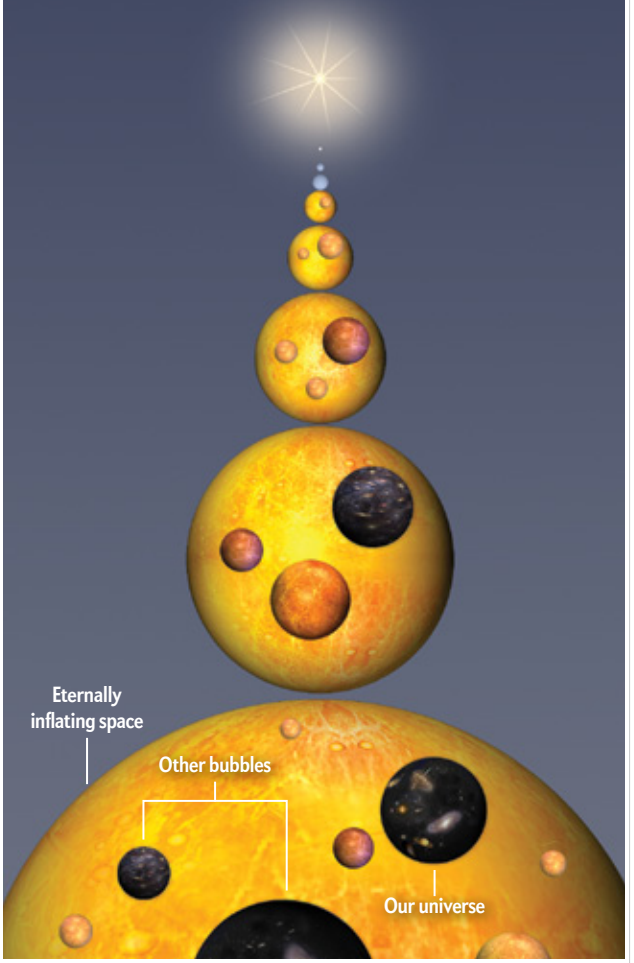
Such procrastinating rogue regions are extremely rare, so you might think it safe to ignore them. You cannot, because they are

© 2014 Scientific American

PROBLEM #3: ETERNAL INFLATION

The Abyss of Infinity

Inflation is known for making precise predictions that have been confirmed by observations. But does it really? Once inflation starts, quantum jittering keeps it going in the bulk of space. Where it does end, a bubble nucleates and grows. We live in such a bubble, but it is atypical; most are younger. In fact, an infinite number of bubbles form with an infinite variety of properties. Everything that can happen happens in some bubble. A theory that predicts everything predicts nothing.



inflating. They continue to grow and, in a matter of instants, dwarf the well-behaved region that ended inflation on time. The result is a sea of inflating space surrounding a little island filled with hot matter and radiation. What is more, rogue regions spawn new rogue regions, as well as new islands of matter—each a self-contained universe. The process continues ad infinitum, creating an unbounded number of islands surrounded by ever more inflating space. If you are not disturbed by this picture, don't worry—you should not be. The disturbing news comes next.

The islands are not all the same. The inherently random nature of quantum physics ensures that some are highly nonuniform or strongly warped. Their nonuniformity sounds like the problem of bad inflation described earlier, but the cause is different. Bad inflation occurs because the parameters controlling the shape of the potential energy curve are likely to be too large. Here nonuniformity can result from eternal inflation and random quantum fluctuations no matter what values the parameters have.

To be quantitatively precise, the word "some" above should be replaced with "an infinite number of." In an eternally inflating universe, an infinite number of islands will have properties like the ones we observe, but an infinite number will not. The true outcome of inflation was best summarized by Guth: "In an eternally inflating universe, anything that can happen will happen; in fact, it will happen an infinite number of times."

So is our universe the exception or the rule? In an infinite collection of islands, it is hard to tell. As an analogy, suppose you have a sack containing a known finite number of quarters and pennies. If you reach in and pick a coin randomly, you can make a firm prediction about which coin you are most likely to choose. If the sack contains an infinite number of quarter and pennies, though, you cannot. To try to assess the probabilities, you sort the coins into piles. You start by putting one quarter into the pile, then one penny, then a second quarter, then a second penny, and so on. This procedure gives you the impression that there is an equal number of each denomination. But then you try a different system, first piling 10 quarters, then one penny, then 10 quarters, then another penny, and so on. Now you have the impression that there are 10 quarters for every penny.

Which method of counting out the coins is right? The answer is neither. For an infinite collection of coins, there are an infinite number of ways of sorting that produce an infinite range of probabilities. So there is no legitimate way to judge which coin is more likely. By the same reasoning, there is no way to judge which kind of island is more likely in an eternally inflating universe.

Now you should be disturbed. What does it mean to say that inflation makes certain predictions—that, for example, the universe has scale-invariant fluctuations—if anything that can happen will happen an infinite number of times? If the theory makes no testable predictions, how can cosmologists claim that the theory agrees with observations, as they routinely do?

THE MEASURE OF OUR FAILURE

THEORISTS ARE NOT UNAWARE of the problem but believe they can somehow restore the naive inflationary picture of the early 1980s. Many remain hopeful even though they have wrestled with this issue for decades without finding a solution.

Some suggest trying to construct theories of inflation that are not eternal, to nip the infinity of universes in the bud. But eternality is a natural consequence of inflation plus quantum physics. To avoid it, the universe would have to start off in a very special initial state and with a special form of inflationary energy, so that inflation ended everywhere in space before quantum fluctuations had a chance to reignite it. In this scenario, though, the observed outcome depends sensitively on what the initial state is. That defeats the entire purpose of inflation: to explain the outcome no matter what conditions existed beforehand.

An alternative strategy supposes that islands like our observable universe are the most likely outcome of inflation. Proponents of this approach impose a so-called measure, a specific rule for weighting which kinds of islands are most likely—analogous to declaring that we must take three quarters for every five pennies when drawing coins from our sack. The notion of a measure, an ad hoc addition, is an open admission that infla-

© 2014 Scientific American

tionary theory on its own does not explain or predict anything.

Worse, theorists have come up with many equally reasonable measures that lead to different conclusions. An example is the volume measure, which says islands should be weighted by size. At first glance, this makes sense. The core idea of inflation is that it explains our universe's uniformity and flatness by creating large volumes of space with those properties. Unfortunately, the volume measure fails because it favors procrastination. Consider two kinds of regions: islands like ours and others that formed later, after more inflation. By the power of exponential growth, the latter regions will occupy vastly more total volume. Hence, regions younger than ours are vastly more common. By this measure, it is unlikely we would even exist.

Measure enthusiasts take a trial-and-error approach in which they invent and test measures until, they hope, one produces the desired answer: that our universe is highly probable. Suppose they succeed someday. Then they will need another principle to justify using that measure instead of the others, yet another principle to choose that principle, and so on.

Still another alternative approach is to invoke the anthropic principle. Whereas the measure concept holds that we live in a typical island, the anthropic principle assumes we live in a very atypical island with just the minimal conditions needed to support life. The claim is that the conditions in more typical islands are incompatible with galaxies or stars or some other prerequisite for life as we know it. Even though the typical islands occupy more space than ones like ours, they can be ignored because we are interested only in regions that humans could potentially inhabit.

But the conditions in our universe are not minimal—the universe is flatter, smoother and more precisely scale-invariant than it had to be to support life. More typical islands, such as those younger than ours, are also habitable yet much more numerous.

MAKING PROCRASTINATORS PAY

IN LIGHT OF THESE ARGUMENTS, the oft-cited claim that cosmological data have verified the central predictions of inflationary theory is misleading, at best. What one can say is that data have confirmed predictions of the naive inflationary theory as we understood it before 1983, but this theory is not inflationary cosmology as understood today. The naive theory supposes that inflation leads to a predictable outcome governed by the laws of classical physics. The truth is that quantum physics rules inflation, and anything that can happen will happen. And if inflationary theory makes no firm predictions, what is its point?

The key problem is that procrastination carries no penalty—to the contrary, it is rewarded. Rogue regions that delay ending inflation continue their accelerating growth, so they invariably take over. Ideally, any rogue regions would expand more slowly—or, better still, shrink. The overwhelming bulk of the universe would consist of well-behaved regions that end the smoothing phase on time, and our observed universe would be comfortably normal.

An alternative to inflationary cosmology that my colleagues and I have proposed, known as the cyclic theory, has just this property. According to this picture, the big bang is not the beginning of space and time [see "The Myth of the Beginning of Time," by Gabriele Veneziano; SCIENTIFIC AMERICAN, May 2004] but rather a "bounce" from a preceding phase of contraction to a new phase of expansion, accompanied by the creation of matter and radiation. The theory is cyclic because, after a trillion years, the expansion devolves into contraction and a new bounce to expansion again. The key point is that the smoothing of the universe occurs before the bang, during the period of contraction. Any procrastinating rogue regions continue to contract while well-behaved regions bounce on time and begin expanding, so the rogue regions remain comparatively small and negligible.

Smoothing during contraction has an observable consequence. During any smoothing phase, whether in inflationary theory or in the cyclic theory, quantum fluctuations generate small, propagating random distortions in spacetime, known as gravitational waves, that leave a distinctive imprint on the microwave background radiation. The amplitude of the waves is proportional to the energy density. Inflation would occur when the universe was extremely dense, whereas the equivalent process in the cyclic model would occur when the universe was practically empty, so the predicted imprints would be very different. Of course, the cyclic theory is relatively new and may have its own problems, but it shows there are alternatives that may not suffer the uncontrollable runaway of eternal inflation. Our work suggests the cyclic model avoids the other problems described earlier, too.

To be sure, I have presented the cases for and against inflation as two extremes without the opportunity for cross-examination or nuance. In the end, the case will be decided by data.

Observations of the microwave background radiation will be telling. In March 2014 researchers using BICEP2, a telescope at the South Pole, announced they had detected the telltale imprint of primordial gravitational waves in the cosmic microwave background. If confirmed, this discovery would help support inflation; if other observations fail to detect the imprint, it would be a major setback for the theory. For inflation to make sense despite a null result, cosmologists would then need to suppose that the inflaton field had a very peculiar potential with just the right shape to suppress gravitational waves, which seems contrived. Many researchers would move to alternatives, such as the cyclic universe theory, that naturally predict an unobservably small gravitational-wave signal. The outcome will be a critical moment in our quest to determine how the universe came to be the way it is and what will happen to it in the future. SA

Paul J. Steinhardt is director of the Princeton Center for Theoretical Science at Princeton University. He is a member of the National Academy of Sciences and received the P.A.M. Dirac Medal from the International Center for Theoretical Physics in 2002 for his contributions to inflationary theory. Steinhardt is also known for postulating a new state of matter known as quasicrystals.

MORE TO EXPLORE

The Inflationary Universe. Alan Guth. Basic Books, 1998.

Quantum Cosmology, Inflation, and the Anthropic Principle. Andrei Linde in *Science and Ultimate Reality: Quantum Theory, Cosmology and Complexity.* Edited by John D. Barrow, Paul C. W. Davies and Charles L. Harper, Jr. Cambridge University Press, 2004.

Endless Universe: Beyond the Big Bang. Paul J. Steinhardt and Neil Turok. Doubleday, 2007.

The Measure Problem in Cosmology. G. W. Gibbons and Neil Turok in *Physical Review D,* Vol. 77, No. 6, Paper No. 063516; March 2008. Preprint online at http://arxiv.org/abs/hep-th/0609095

From Eternity to Here: The Quest for the Ultimate Theory of Time. Sean Carroll. Dutton Adult, 2010.

scientificamerican.com/magazine/sa

© 2014 Scientific American

PHILOSOPHY OF SCIENCE

DOES THE MULTIVERSE REALLY EXIST?

Proof of parallel universes radically different from our own may still lie beyond the domain of science

By George F. R. Ellis

IN THE PAST DECADE AN EXTRAORDINARY CLAIM HAS CAPTIVATED COSMOLOGISTS: that the expanding universe we see around us is not the only one; that billions of other universes are out there, too. There is not one universe—there is a multiverse. In *Scientific American* articles and books such as Brian Greene's *The Hidden Reality,* leading scientists have spoken of a super-Copernican revolution. In this view, not only is our planet one among many, but even our entire universe is insignificant on the cosmic scale of things. It is just one of countless universes, each doing its own thing. The word "multiverse" has different meanings. Astronomers are able to see out to a distance of about 42 billion light-years, our cosmic visual horizon. We have no reason to suspect the universe stops there. Beyond it could be many—even infinitely many—domains much like the one we see. Each has a different initial distribution of matter, but the same laws of physics operate in all. Nearly all cosmologists today (including me) accept this type of multiverse, which Max Tegmark calls "level 1." Yet some go further. They suggest completely different kinds of universes, with different physics, different histories, maybe different numbers of spatial dimensions. Most will be sterile, although some will be teeming with life. A chief proponent of this "level 2" multiverse is Alexander Vilenkin, who paints a dramatic picture of an infinite set of universes with an infinite number of galaxies,

IN BRIEF

The notion of parallel universes leapt out of the pages of fiction into scientific journals in the 1990s. Many scientists claim that mega-millions of other universes, each with its own laws of physics, lie out there, beyond our visual horizon. They are collectively known as the multiverse. **The trouble** is that no possible astronomical observations can ever see those other universes. The arguments are indirect at best. And even if the multiverse exists, it leaves the deep mysteries of nature unexplained.

Photograph by Levi Brown

© 2014 Scientific American

© 2014 Scientific American

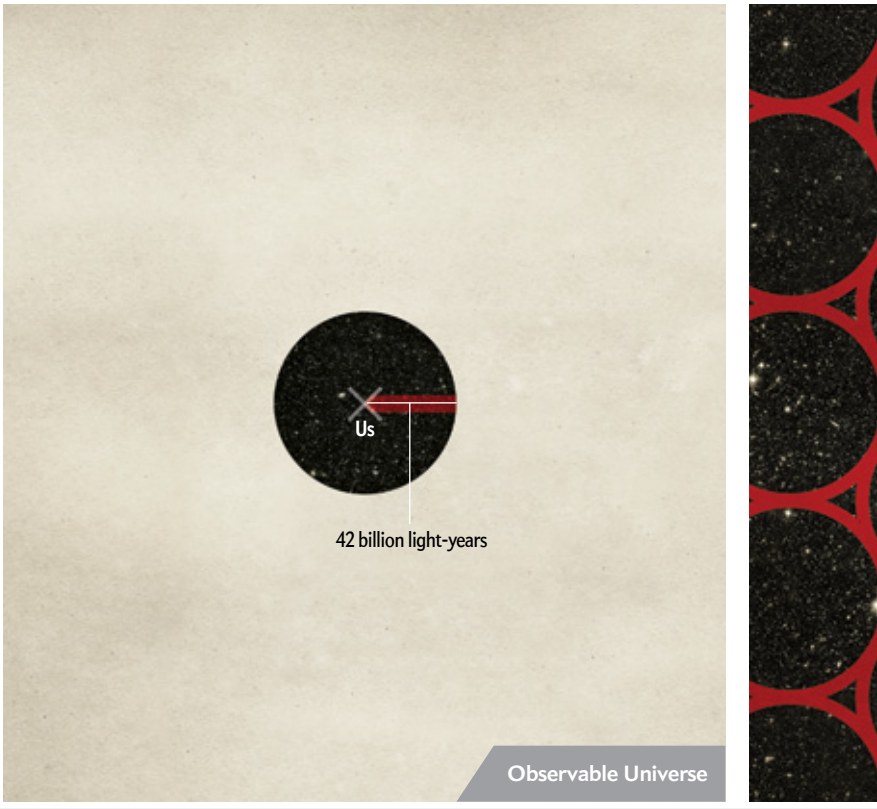
THE PERILS OF EXTRAPOLATION

What Lies Beyond?

When astronomers peer into the universe, they see out to a distance of about 42 billion light-years, our cosmic horizon, which represents how far light has been able to travel since the big bang (as well as how much the universe has expanded in size since then). Assuming that space does not just stop there and may well be infinitely big, cosmologists make educated guesses as to what the rest of it looks like.

Level 1 Multiverse: Plausible The most straightforward assumption is that our volume of space is a representative sample of the whole. Distant alien beings see different volumes, but all of these look basically alike, apart from random variations in the distribution of matter. Together these regions, seen and unseen, form a basic type of multiverse.

Level 2 Multiverse: Questionable Many cosmologists go further and speculate that, sufficiently far away, things look quite different from what we see. Our environs may be one of many bubbles floating in an otherwise empty background. The laws of physics would differ from bubble to bubble, leading to an almost inconceivable variety of outcomes. Those other bubbles may be impossible to observe even in principle. The author and other skeptics feel dubious about this type of multiverse.



an infinite number of planets and an infinite number of people with your name who are reading this article.

Similar claims have been made since antiquity by many cultures. What is new is the assertion that the multiverse is a scientific theory, with all that implies about being mathematically rigorous and experimentally testable. I am skeptical about this claim. I do not believe the existence of those other universes has been proved—or ever could be. Proponents of the multiverse, as well as greatly enlarging our conception of physical reality, are implicitly redefining what is meant by "science."

OVER THE HORIZON

THOSE WHO SUBSCRIBE to a broad conception of the multiverse have various proposals as to how such a proliferation of universes might arise and where they would all reside. They might be sitting in regions of space far beyond our own, as envisaged by the chaotic inflation model of Alan H. Guth, Andrei Linde and others [see "The Self-Reproducing Inflationary Universe," by Andrei Linde; SCIENTIFIC AMERICAN, November 1994]. They might exist at different epochs of time, as proposed in the cyclic universe model of Paul J. Steinhardt and Neil Turok [see "The Myth of the Beginning of Time," by Gabriele Veneziano; Scientific American, May 2004]. They might exist in the same space we do but in a different branch of the quantum wave function, as advocated by David Deutsch [see "The Quantum Physics of Time Travel," by David Deutsch and Michael Lockwood; SCIENTIFIC AMERICAN, March 1994]. They might not have a location, being completely disconnected from our spacetime, as suggested by Tegmark and Dennis Sciama [see "Parallel Universes," by Max Tegmark; SCIENTIFIC AMERICAN, May 2003].

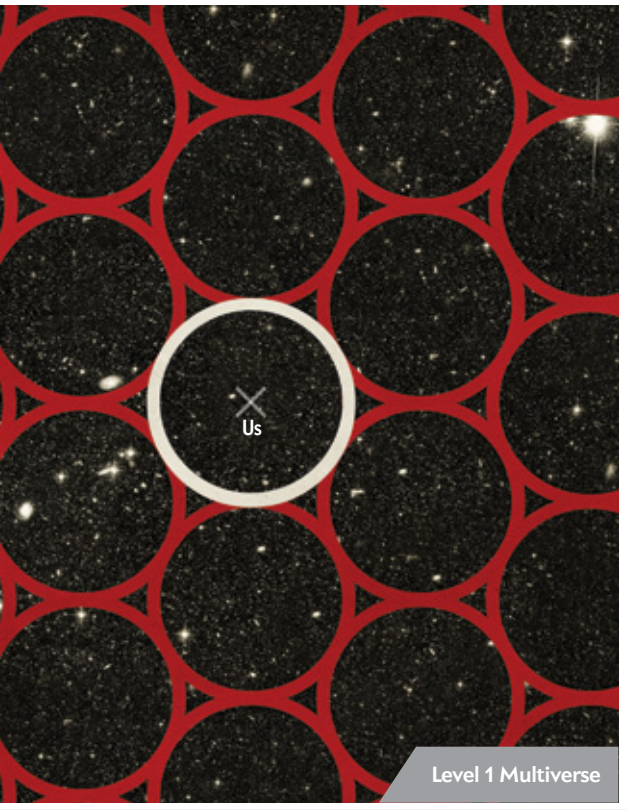
Of these options, the most widely accepted is that of chaotic inflation, and I will concentrate on it; however, most of my remarks apply to all the other proposals as well. The idea is that space at large is an eternally expanding void, within which quantum effects continually spawn new universes like a child blowing bubbles. The concept of inflation goes back to the 1980s, and physicists have elaborated on it based on their most comprehensive theory of nature: string theory. String theory allows bubbles to look very different from one another. In effect, each begins life not only with a random distribution of matter but also with random types of matter. Our universe contains particles such as electrons and quarks interacting through forces such as electromagnetism; other universes may have very different types of particles and forces—which is to say, different local laws of physics. The full set of allowed local laws is known as the landscape. In some interpretations of string theory, the landscape is immense, ensuring a tremendous diversity of universes.

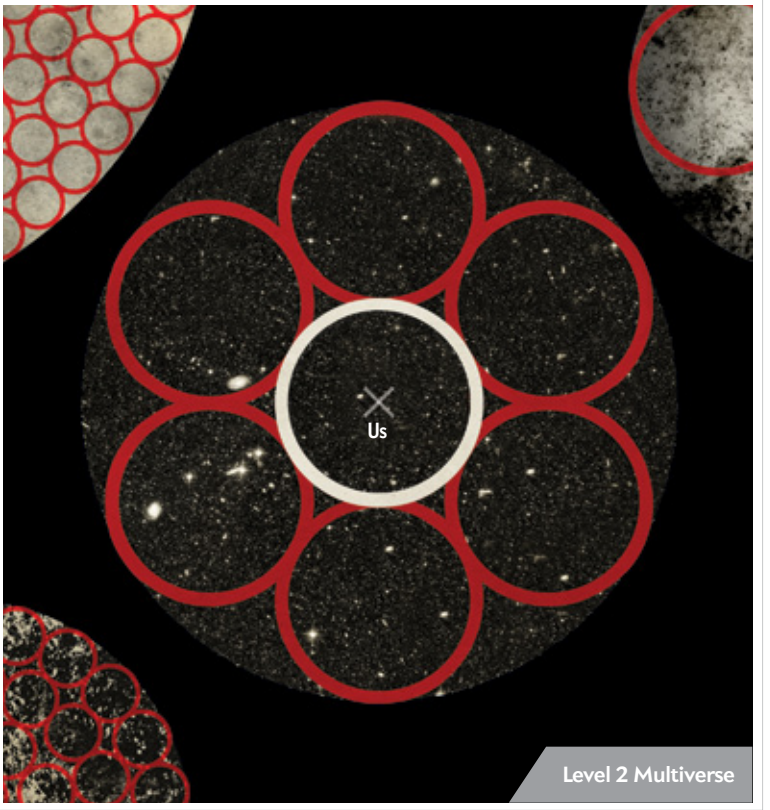
Many physicists who talk about the multiverse, especially advocates of the string landscape, do not care much about parallel universes per se. For them, objections to the multiverse as a concept are unimportant. Their theories live or die based on internal consistency and, one hopes, eventual laboratory testing. They assume a multiverse context for their theories without worrying about how it comes to be—which is what concerns cosmologists.

For a cosmologist, the basic problem with all multiverse proposals is the presence of a cosmic visual horizon. The horizon is the limit to how far away we can see, because signals traveling toward us at the speed of light (which is finite) have not had time since the beginning of the universe to reach us from farther out. All the parallel universes lie outside our horizon and remain beyond our capacity to see, now or ever, no matter how technol-

Illustrations by Chad Hagen

© 2014 Scientific American





ogy evolves. In fact, they are too far away to have had any influence on our universe whatsoever. That is why none of the claims made by multiverse enthusiasts can be directly substantiated.

The proponents are telling us we can state in broad terms what happens 1,000 times as far as our cosmic horizon, 10^{100} times, $10^{1,000,000}$ times, an infinity—all from data we obtain within the horizon. It is an extrapolation of an extraordinary kind. Maybe the universe closes up on a very large scale, and there is no infinity out there. Maybe all the matter in the universe ends somewhere, and there is empty space forever after. Maybe space and time come to an end at a singularity that bounds the universe. We just do not know what actually happens, for we have no information about these regions and never will.

SEVEN QUESTIONABLE ARGUMENTS

MOST MULTIVERSE PROPONENTS are careful scientists who are quite aware of this problem but think we can still make educated guesses about what is going on out there. Their arguments fall into seven broad types, each of which runs into trouble.

Space has no end. Few dispute that space extends beyond our cosmic horizon and that many other domains lie beyond what we see. If this limited type of multiverse exists, we can extrapolate what we see to domains beyond the horizon, with more and more uncertainty as regards the farther-out regions. It is then easy to imagine more elaborate types of variation, including alternative physics occurring out where we cannot see. But the trouble with this type of extrapolation, from the known to the unknown, is that no one can prove you wrong. How can scientists decide whether their picture of an unobservable region of spacetime is a reasonable or an unreasonable extrapolation of what we see? Might other universes have different initial distributions of matter, or might they also have different values of fundamental physical constants, such as those that set the strength of nuclear forces? You could get either, depending on what you assume.

Known physics predicts other domains. Proposed unified theories predict entities such as scalar fields, a hypothesized relative of other space-filling fields such as the magnetic field. Such fields should drive cosmic inflation and create universes ad infinitum. These theories are well grounded theoretically, but the nature of the hypothesized fields is unknown, and experimentalists have yet to demonstrate their existence, let alone measure their supposed properties. Crucially, physicists have not substantiated that the dynamics of these fields would cause different laws of physics to operate in different bubble universes.

The theory that predicts an infinity of universes passes a key observational test. The cosmic microwave background radiation reveals what the universe looked like at the end of its hot early expansion era. Patterns in it suggest that our universe really did undergo a period of inflation. But not all types of inflation go on forever and create an infinite number of bubble universes. Observations do not single out the required type of inflation from other types. Some cosmologists such as Steinhardt even argue that eternal inflation would have led to different patterns in the background radiation than we see [see "The Inflation Debate," by Paul J. Steinhardt; SCIENTIFIC AMERICAN, April 2011]. Linde and others disagree. Who is right? It all depends on what you assume about the physics of the inflationary field.

Fundamental constants are finely tuned for life. A remarkable fact about our universe is that physical constants have just the right values needed to allow for complex structures, in-

© 2014 Scientific American

cluding living things. Steven Weinberg, Martin Rees, Leonard Susskind and others contend that an exotic multiverse provides a tidy explanation for this apparent coincidence: if all possible values occur in a large enough collection of universes, then viable ones for life will surely be found somewhere. This reasoning has been applied, in particular, to explaining the density of the dark energy that is speeding up the expansion of the universe today. I agree that the multiverse is a possible valid explanation for the value of this density; arguably, it is the only scientifically based option we have right now. But we have no hope of testing it observationally. Additionally, most analyses of the issue assume the basic equations of physics are the same everywhere, with only the constants differing—but if one takes the multiverse seriously, this need not be so [see "Looking for Life in the Multiverse," by Alejandro Jenkins and Gilad Perez; SCIENTIFIC AMERICAN, January 2010].

Fundamental constants match multiverse predictions. This argument refines the previous one by suggesting that the universe is no more finely tuned for life than it strictly needs to be. Proponents have assessed the probabilities of various values of the dark energy density. The higher the value is, the more probable it is, but the more hostile the universe would be to life. The value we observe should be just on the borderline of uninhabitability, and it does appear to be so [*see box at right*]. Where the argument stumbles is that we cannot apply a probability argument if there is no multiverse to apply the concept of probability to. This argument thus assumes the desired outcome before it starts; it simply is not applicable if there is only one physically existing universe. Probability is a probe of the consistency of the multiverse proposal, not a proof of its existence.

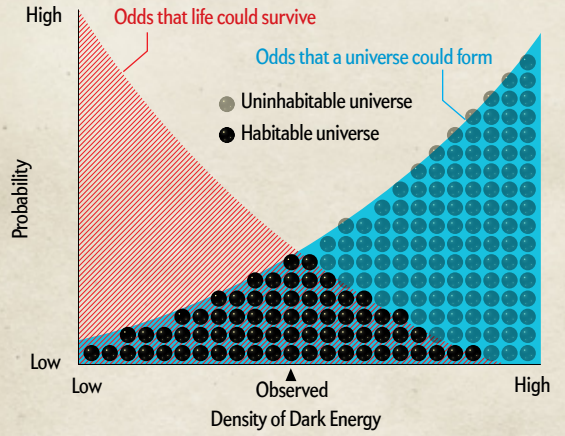
String theory predicts a diversity of universes. String theory has moved from being a theory that explains everything to a theory where almost anything is possible. In its current form, it predicts that many essential properties of our universe are pure happenstance. If the universe is one of a kind, those properties seem inexplicable. How can we understand, for example, the fact that physics has precisely those highly constrained properties that allow life to exist? If the universe is one of many, those properties make perfect sense. Nothing singled them out; they are simply the ones that arose in our region of space. Had we lived elsewhere, we would have observed different properties, if we could indeed exist there (life would be impossible in most places). But string theory is not a tried-and-tested theory; it is not even a complete theory. If we had proof that string theory is correct, its theoretical predictions could be a legitimate, experimentally based argument for a multiverse. We do not have such proof.

All that can happen, happens. In seeking to explain why nature obeys certain laws and not others, some physicists and philosophers have speculated that nature never made any such choice: all conceivable laws apply somewhere. The idea is inspired in part by quantum mechanics, which, as Murray Gell-Mann memorably put it, holds that everything not forbidden is compulsory. A particle takes all the paths it can, and what we see is the weighted average of all those possibilities. Perhaps the same is true of the entire universe, implying a multiverse. But astronomers have not the slightest chance of observing this multiplicity of possibilities. Indeed, we cannot even know what the possibilities are. We can only make sense of this proposal in the face of some unverifiable organizing principle or framework that decides what is allowed and what is not—for example, that all possible mathematical structures must be realized in some physical domain (as proposed by Tegmark). But we have no idea what kinds of existence this principle entails, apart from the fact that it must, of necessity, include the world we see around us. And we have no way whatsoever to verify the existence or nature of any such organizing principle. It is in some ways an attractive proposition, but its proposed application to reality is pure speculation.

DARK ENERGY AND THE MULTIVERSE

Does the Glove Fit?

As evidence for a multiverse, proponents often cite the density of the dark energy that dominates our universe. The process of eternal inflation endows each universe in a multiverse with a random density of dark energy. Relatively few universes have zero or a low value; most have higher values (*blue area*). But too much dark energy tears apart the complex structures needed to sustain life (*red area*). So most habitable universes should have a middling density of dark energy (*peak of overlap region*)—and, lo and behold, our universe does. Multiverse skeptics, though, say this reasoning is circular: it holds only if you assume the multiverse to begin with. It is a consistency test, not a proof.



ABSENCE OF EVIDENCE

ALTHOUGH THE THEORETICAL ARGUMENTS fall short, cosmologists have also suggested various empirical tests for parallel universes. The cosmic microwave background radiation might bear some traces of other bubble universes if, for example, our universe has ever collided with another bubble of the kind implied by the chaotic inflation scenario. The background radiation might also contain remnants of universes that existed before the big bang in an endless cycle of universes. These are indeed ways one might get real evidence of other universes. Some cosmologists have even claimed to see such remnants. The observational claims are strongly disputed, however, and many of the hypothetically possible multiverses would not lead to such evidence. So observers can test only some specific classes of multiverse models in this way.

A second observational test is to look for variations in one or more fundamental constants, which would corroborate the premise that the laws of physics are not so immutable after all. Some astronomers claim to have found such variations [see "Inconstant

Graphic by Jen Christiansen

© 2014 Scientific American

Constants," by John D. Barrow and John K. Webb; Scientific American, June 2005]. Most, though, consider the evidence dubious.

A third test is to measure the shape of the observable universe: Is it spherical (positively curved), hyperbolic (negatively curved) or "flat" (uncurved)? Multiverse scenarios generally predict that the universe is not spherical, because a sphere closes up on itself, allowing for only a finite volume. Unfortunately, this test is not a clean one. The universe beyond our horizon could have a different shape from that in the observed part; what is more, not all multiverse theories rule out a spherical geometry.

A better test is the topology of the universe: Does it wrap around like a doughnut or pretzel? If so, it would be finite in size, which would definitely disprove multiverse scenarios based on chaotic inflation. Such a shape would produce recurring patterns in the sky, such as giant circles in the cosmic microwave background radiation. Observers have looked for and failed to find any such patterns. But this null result cannot be taken as a point in favor of the multiverse. It just means that specific types of a single universe are ruled out.

Finally, physicists might hope to prove or disprove some of the theories that predict a multiverse. They might find observational evidence against chaotic versions of inflation or discover a mathematical or empirical inconsistency that forces them to abandon the landscape of string theory. That scenario would undermine much of the motivation for supporting the multiverse idea, although it would not rule the concept out altogether.

TOO MUCH WIGGLE ROOM

All in all, the case for the multiverse is inconclusive. The basic reason is the extreme flexibility of the proposal: it is more a concept than a well-defined theory. Most proposals involve a patchwork of different ideas rather than a coherent whole. The basic mechanism for eternal inflation does not itself cause physics to be different in each domain in a multiverse; for that, it needs to be coupled to another speculative theory. Although they can be fitted together, there is nothing inevitable about it.

The key step in justifying a multiverse is extrapolation from the known to the unknown, from the testable to the untestable. You get different answers depending on what you choose to extrapolate. Because theories involving a multiverse can explain almost anything whatsoever, any observation can be accommodated by some multiverse variant. The various "proofs," in effect, propose that we should accept a theoretical explanation instead of insisting on observational testing. But such testing has, up until now, been the central requirement of the scientific endeavor, and we abandon it at our peril. If we weaken the requirement of solid data, we weaken the core reason for the success of science over the past centuries.

Now, it is true that a unifying explanation of some range of phenomena carries greater weight than a hodgepodge of separate arguments for the same phenomena. If the unifying explanation necessarily assumes the existence of unobservable entities such as parallel universes, we might feel compelled to accept those entities. But a key issue here is how many unverifiable entities are needed. Specifically, are we hypothesizing more or fewer entities than the number of phenomena to be explained? In the case of the multiverse, we are supposing the existence of a huge number—perhaps even an infinity—of unobservable entities to explain just one existing universe. It hardly fits 14th-century English philosopher William of Ockham's stricture that "entities must not be multiplied beyond necessity."

Proponents of the multiverse make one final argument: that there are no good alternatives. As distasteful as scientists might find the proliferation of parallel worlds, if it is the best explanation, we would be driven to accept it; conversely, if we are to give up the multiverse, we need a viable alternative. This exploration of alternatives depends on what kind of explanation we are prepared to accept. Physicists' hope has always been that the laws of nature are inevitable—that things are the way they are because there is no other way they might have been—but we have been unable to show this is true. Other options exist, too. The universe might be pure happenstance—it just turned out that way. Or things might in some sense be meant to be the way they are—purpose or intent somehow underlies existence. Science cannot determine which is the case, because these are metaphysical issues.

Scientists proposed the multiverse as a way of resolving deep issues about the nature of existence, but the proposal leaves the ultimate issues unresolved. All the same issues that arise in relation to the universe arise again in relation to the multiverse. If the multiverse exists, did it come into existence through necessity, chance or purpose? That is a metaphysical question that no physical theory can answer for either the universe or the multiverse.

To make progress, we need to keep to the idea that empirical testing is the core of science. We need some kind of causal contact with whatever entities we propose; otherwise, there are no limits. The link can be a bit indirect. If an entity is unobservable but absolutely essential for properties of other entities that are indeed verified, it can be taken as verified. But then the onus of proof devolves to showing that it is absolutely essential to the web of explanation. The challenge I pose to multiverse proponents is: Can you prove that unseeable parallel universes are vital to explain the world we do see? And is the link essential and inescapable?

As skeptical as I am, I think the contemplation of the multiverse is an excellent opportunity to reflect on the nature of science and on the ultimate nature of existence: why we are here. It leads to new and interesting insights and so is a productive research program. In looking at this concept, we need an open mind, though not too open. It is a delicate path to tread. Parallel universes may or may not exist; the case is unproved. We are going to have to live with that uncertainty. Nothing is wrong with scientifically based philosophical speculation, which is what multiverse proposals are. But we should name it for what it is. SA

George F. R. Ellis is a cosmologist and emeritus mathematics professor at the University of Cape Town in South Africa. He is one of the world's leading experts on Einstein's general theory of relativity and co-author, with Stephen Hawking, of the seminal book *The Large Scale Structure of Space-Time* (Cambridge University Press, 1973).

MORE TO EXPLORE

Issues in the Philosophy of Cosmology. George F. R. Ellis in *Philosophy of Physics.* Edited by Jeremy Butterfield and John Earman. Elsevier, 2006. arxiv.org/abs/astro-ph/0602280

Universe or Multiverse? Edited by Bernard Carr. Cambridge University Press, 2009.

The Hidden Reality: Parallel Universes and the Deep Laws of the Cosmos. Brian Greene. Knopf, 2011.

Higher Speculations: Grand Theories and Failed Revolutions in Physics and Cosmology. Helge Kragh. Oxford University Press, 2011.

scientificamerican.com/magazine/sa

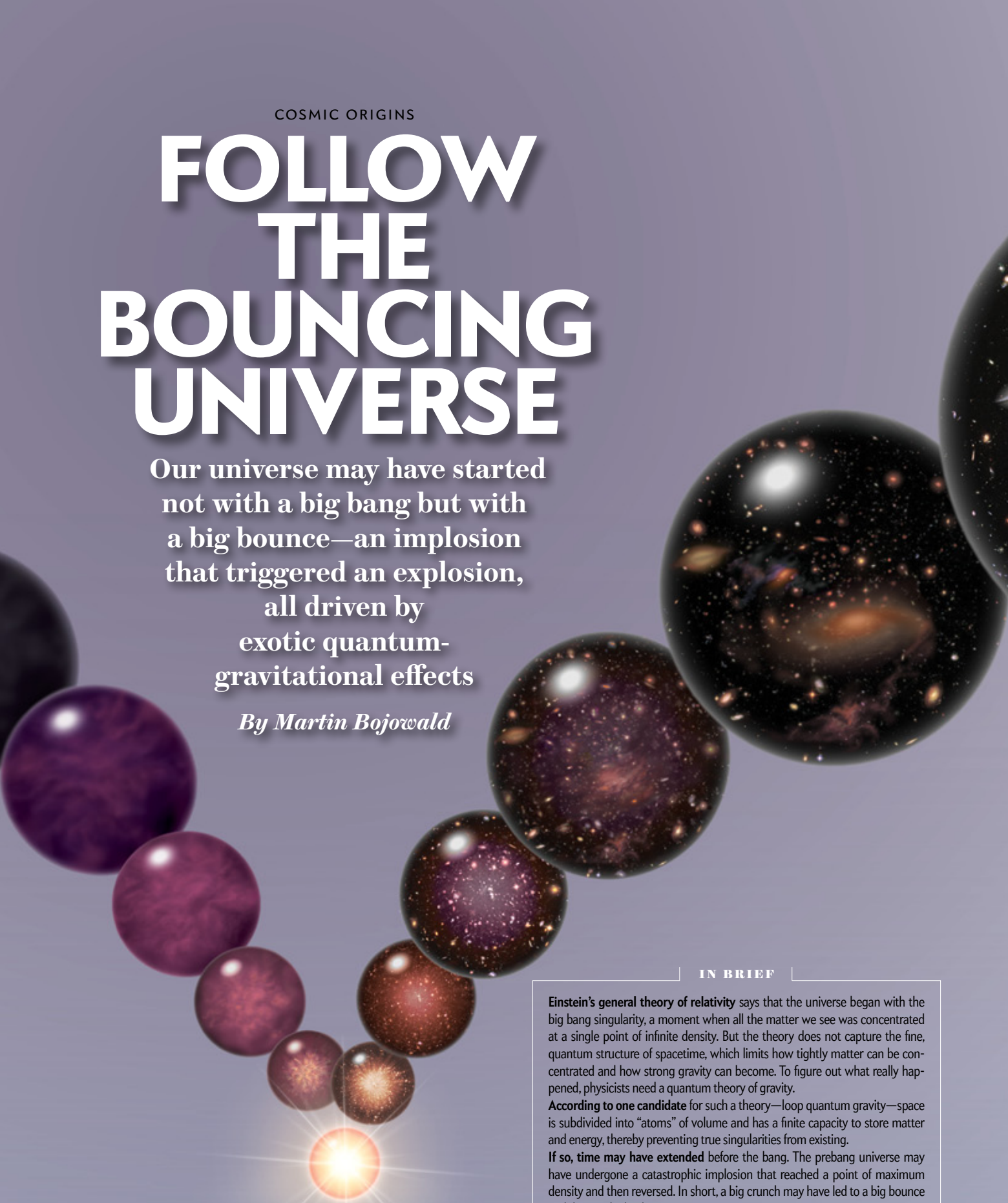
© 2014 Scientific American

COSMIC ORIGINS

FOLLOW THE BOUNCING UNIVERSE

Our universe may have started not with a big bang but with a big bounce—an implosion that triggered an explosion, all driven by exotic quantum-gravitational effects

By Martin Bojowald



IN BRIEF

Einstein's general theory of relativity says that the universe began with the big bang singularity, a moment when all the matter we see was concentrated at a single point of infinite density. But the theory does not capture the fine, quantum structure of spacetime, which limits how tightly matter can be concentrated and how strong gravity can become. To figure out what really happened, physicists need a quantum theory of gravity.

According to one candidate for such a theory—loop quantum gravity—space is subdivided into "atoms" of volume and has a finite capacity to store matter and energy, thereby preventing true singularities from existing.

If so, time may have extended before the bang. The prebang universe may have undergone a catastrophic implosion that reached a point of maximum density and then reversed. In short, a big crunch may have led to a big bounce and then to the big bang.

Illustration by Pat Rawlings, SAIC

© 2014 Scientific American

ATOMS ARE NOW SUCH A COMMONplace idea that it is hard to remember how radical they used to seem. When scientists first hypothesized atoms centuries ago, they despaired of ever observing anything so small, and many questioned whether the concept of atoms could even be called scientific. Gradually, however, evidence for atoms accumulated and reached a tipping point with Albert Einstein's 1905 analysis of Brownian motion, the random jittering of dust grains in a fluid. Even then, it took another 20 years for physicists to develop a theory explaining atoms—namely, quantum mechanics—and another 30 for physicist Erwin Müller to make the first microscope images of them. Today entire industries are based on the characteristic properties of atomic matter.

Physicists' understanding of the composition of space and time is following a similar path, but several steps behind. Just as the behavior of materials indicates that they consist of atoms, the behavior of space and time suggests that they, too, have some fine-scale structure—either a mosaic of spacetime "atoms" or some other filigree work. Material atoms are the smallest indivisible units of chemical compounds; similarly, the putative space atoms are the smallest indivisible units of distance. They are generally thought to be about 10^{-35} meter in size, far too tiny to be seen by today's most powerful instruments, which probe distances as short as 10^{-20} meter. Consequently, many scientists question whether the concept of atomic spacetime can even be called scientific. Undeterred, other researchers are coming up with possible ways to detect such atoms indirectly.

The most promising involve observations of the cosmos. If we imagine rewinding the expansion of the universe back in time, the

© 2014 Scientific American

galaxies we see all seem to converge on a single infinitesimal point: the big bang singularity. At this point, our current theory of gravity—Einstein's general theory of relativity—predicts that the universe had an infinite density and temperature. This moment is sometimes sold as the beginning of the universe, the birth of matter, space and time. Such an interpretation, however, goes too far because the infinite values indicate that general relativity itself breaks down. To explain what really happened at the big bang, physicists must transcend relativity. We must develop a theory of quantum gravity, which would capture the fine structure of spacetime to which relativity is blind.

The details of that structure came into play under the dense conditions of the primordial universe, and traces of it may survive in the present-day arrangement of matter and radiation. In short, if spacetime atoms exist, it will not take centuries to find the evidence, as it did for material atoms. With some luck, we may know within this decade.

PIECES OF SPACE

PHYSICISTS HAVE DEVISED several candidate theories of quantum gravity, each applying quantum principles to general relativity in a distinct way. My work focuses on the theory of loop quantum gravity ("loop gravity," for short), which was developed in the 1990s using a two-step procedure. First, theorists mathematically reformulated general relativity to resemble the classical theory of electromagnetism; the eponymous "loops" of the theory are analogues of electric and magnetic field lines. Second, following innovative procedures, some that are akin to the mathematics of knots, they applied quantum principles to the loops. The resulting quantum gravity theory predicts the existence of spacetime atoms [see "Atoms of Space and Time," by Lee Smolin; SCIENTIFIC AMERICAN, January 2004].

Other approaches, such as string theory and so-called causal dynamical triangulations, do not predict spacetime atoms per se but suggest other ways that sufficiently short distances might be indivisible [see "The Great Cosmic Roller-Coaster Ride," by Cliff Burgess and Fernando Quevedo; SCIENTIFIC AMERICAN, November 2007, and "The Self-Organizing Quantum Universe," by Jan Ambjørn, Jerzy Jurkiewicz and Renate Loll; SCIENTIFIC AMERICAN, July 2008]. The differences among these theories have given rise to controversy, but to my mind the theories are not contradictory so much as complementary. String theory, for example, is useful for a unified view of particle interactions, including gravity when it is weak. For the purpose of disentangling what happens at the singularity, where gravity is strong, the atomic constructions of loop gravity are more useful.

The theory's power is its ability to capture the fluidity of spacetime. Einstein's great insight was that spacetime is no mere stage on which the drama of the universe unfolds. It is an actor in its own right. It not only determines the motion of bodies within the universe, but it evolves. A complicated interplay between matter and spacetime ensues. Space can grow and shrink.

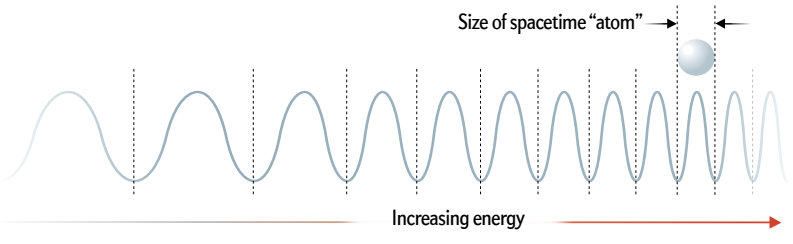
Loop gravity extends this insight into the quantum realm. It takes our familiar understanding of particles of matter and applies it to the atoms of space and time, providing a unified view of our most basic concepts. For instance, the quantum theory of electromagnetism describes a vacuum devoid of particles such as photons, and each increment of energy added to this vacuum generates a new particle. In the quantum theory of gravity, a vacuum is the absence of spacetime—an emptiness so thorough we can scarcely imagine it. Loop gravity describes how each increment of energy added to this vacuum generates a new atom of spacetime.

The spacetime atoms form a dense, ever shifting mesh. Over

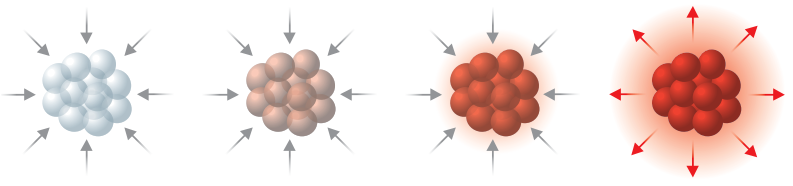
WHAT SPACETIME ATOMS DO

Looking Quite Repulsive

As you increase the energy of radiation in a volume of space, its wavelength approaches that of spacetime "atoms." The space becomes saturated like a waterlogged sponge.



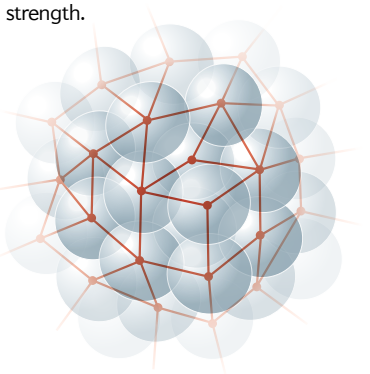
This means space can run out of room to store energy. If you try to pack in still more energy, space will push it back out, creating a repulsive gravitational force. This behavior would preclude the possibility of a big bang singularity.



LOOP QUANTUM GRAVITY

Atoms of Space

Relativity theory runs into trouble because it assumes space is a continuum. A more sophisticated theory, such as loop quantum gravity, holds that space is a mesh of tiny "atoms" (*spheres*). The diameter of these atoms (*red lines*) is the so-called Planck length, the distance over which gravitational and quantum effects are comparable in strength.



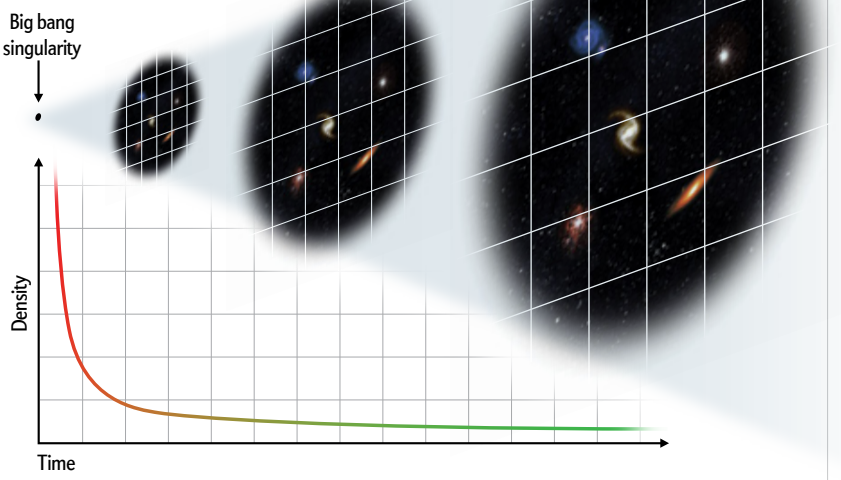
Illustrations by Samuel Velasco, 5W Infographics

© 2014 Scientific American

THE PROBLEM

The Theory Just Went Bang

The idea of the big bang comes from a simple observed fact: galaxies in the universe are moving apart. If you play this trend back in time, galaxies (or their precursors) must have been all scrunched up 13.8 billion years ago. In fact, according to Einstein's general theory of relativity, before the universe began expanding, they were scrunched into a single point of infinite density—the big bang singularity. But an infinite density is unrealistic: that relativity theory predicts it is a sign that the theory is incomplete.



large distances, their dynamism gives rise to the evolving universe of classical general relativity. Under ordinary conditions, we never notice the existence of these spacetime atoms; the mesh spacing is so tight that it looks like a continuum. But when spacetime is packed with energy, as it was at the big bang, the fine structure of spacetime becomes a factor, and the predictions of loop gravity diverge from those of general relativity.

ATTRACTED TO REPULSION

APPLYING THE THEORY is an extremely complex task, so my colleagues and I use simplified versions that capture the truly essential features of the universe, such as its size, and ignore details of lesser interest. We have also had to adapt many of the standard mathematical tools of physics and cosmology. For instance, theoretical physicists commonly describe the world using differential equations, which specify the rate of change of physical variables, such as density, at each point in the spacetime continuum. But when spacetime is grainy, we instead use so-called difference equations, which break up the continuum into discrete intervals. These equations describe how a universe climbs up the ladder of sizes that it is allowed to take as it grows. When I set out to analyze the cosmological implications of loop gravity in 1999, most researchers expected that these difference equations would simply reproduce old results in disguise. But unexpected features soon emerged.

Gravity is typically an attractive force. A ball of matter tends to collapse under its own weight, and if its mass is sufficiently large, gravity overpowers all other forces and compresses the ball into a singularity, such as the one at the center of a black hole. But loop gravity suggests that the atomic structure of spacetime changes the nature of gravity at very high energy densities, making it repulsive. Imagine space as a sponge and mass and energy as water. The porous sponge can store water but only up to a certain amount. Fully soaked, it can absorb no more and instead repels water. Similarly, an atomic quantum space is porous and has a finite amount of storage space for energy. When energy densities become too large, repulsive forces come into play. The continuous space of general relativity, in contrast, can store a limitless amount of energy.

Because of the quantum-gravitational change in the balance of forces, no singularity—no state of infinite density—can ever arise. According to this model, matter in the early universe had a very high but finite density, the equivalent of a trillion suns in every proton-size region. At such extremes, gravity acted as a repulsive force, causing space to expand; as densities moderated, gravity switched to being the attractive force we all know. Inertia has kept the expansion going to the present day.

In fact, the repulsive gravity caused space to expand at an accelerating rate. Cosmological observations appear to require such an early period of acceleration, known as cosmic inflation. As the universe expands, the force driving inflation slowly subsides. Once the acceleration ends, surplus energy is transferred to ordinary matter, which begins to fill the universe in a process called reheating. In current models, inflation is somewhat ad hoc—added in to conform to observations—but in loop quantum cosmology, it is a natural consequence of the atomic nature of spacetime. Acceleration automatically occurs when the universe is small and its porous nature still quite significant.

TIME BEFORE TIME

WITHOUT A SINGULARITY to demarcate the beginning of time, the history of the universe may extend further back than cosmologists once thought possible. Other physicists have reached a similar conclusion [see "The Myth of the Beginning of Time," by Gabriele Veneziano; SCIENTIFIC AMERICAN, May 2004], but only rarely do their models fully resolve the singularity; most models, including those from string theory, require assumptions as to what might have happened at this uneasy spot. Loop gravity, in contrast, is able to trace what took place at the singularity. Loop-based scenarios, though admittedly simplified, are founded on general principles and avoid introducing new ad hoc assumptions.

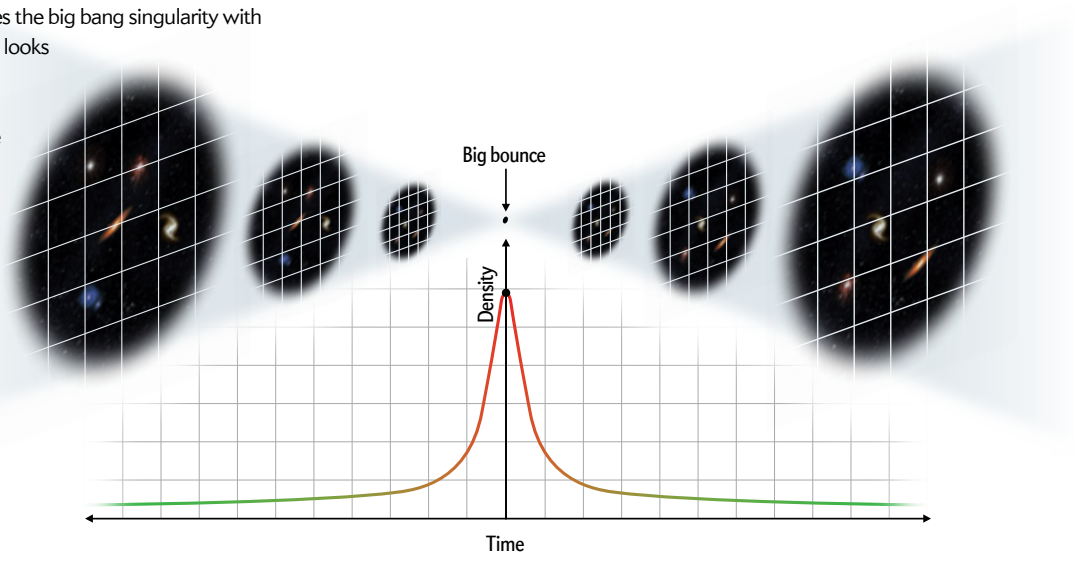
Using the difference equations, we can try to reconstruct the deep past. One possible scenario is that the initial high-density

© 2014 Scientific American

NEW VIEW OF COSMIC ORIGINS

Replacing the Bang

By setting a limit to how much energy you can pack into space, loop quantum gravity replaces the big bang singularity with a big bounce—a process that looks like a beginning but actually reflects a transition from a preexisting state. The bounce sets the expansion of the universe in motion.



In one scenario, the universe is eternal. It imploded to maximal density, which caused gravity to turn repulsive (the bounce), and blew apart again.

THE UNIVERSE

BEFORE THE BOUNCE

Mirror, Mirror

Despite the effects that scrambled the universe during the big bounce, physicists can make some educated guesses about what came before. Some are truly strange. For instance, the difference equations of loop quantum gravity imply that the region of spacetime preceding the bounce was a mirror image of space in our universe. That is, what was right-handed after the bang was left-handed before, and vice versa.

To visualize this effect, imagine a deflating balloon that, instead of settling down to a limp piece of rubber, preserves its energy and momentum. The rubber, once set in motion, tends to stay in motion. So, as the balloon collapses down to a minimum size, it turns inside out and starts growing again. What was previously the outside of the balloon becomes the inside, and vice versa. Similarly, when spacetime atoms cross one another at the big bounce, the universe turns inside out.

This reversal is interesting because elementary particles are not perfectly mirror-symmetric; certain processes change when the orientation does. This asymmetry has to be taken into account to understand what happens to matter at the bounce.

—*M.B.*

state arose when a preexisting universe collapsed under the attractive force of gravity. The density grew so high that gravity switched to being repulsive, and the universe started expanding again. Cosmologists refer to this process as a bounce.

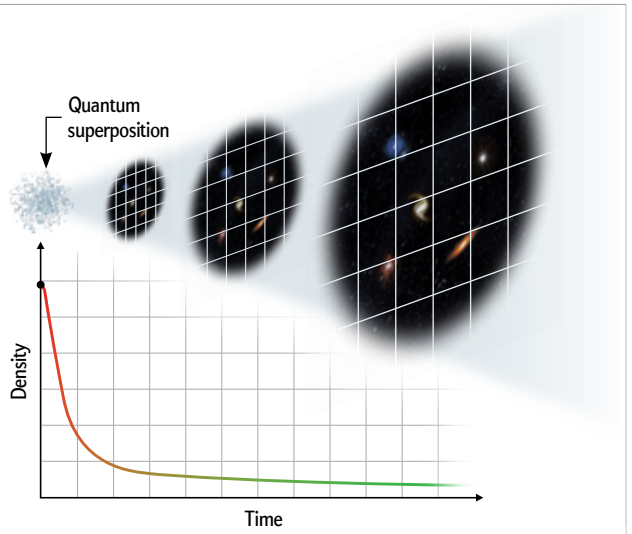
The first loopy bounce model investigated thoroughly was an idealized case in which the universe was highly symmetrical and contained just one type of matter. Particles had no mass and did not interact with one another. Simplified though this model was, understanding it required a set of numerical simulations that were completed only in 2006 by Abhay Ashtekar, Tomasz Pawlowski and Parampreet Singh, all then at Pennsylvania State University. They considered the propagation of waves representing the universe both before and after the big bang. The model clearly showed that a wave would not blindly follow the classical trajectory into the abyss of a singularity but would stop and turn back once the repulsion of quantum gravity set in.

An exciting result of these simulations was that the notorious uncertainty of quantum mechanics seemed to remain fairly muted during the bounce. A wave remained localized throughout the bounce rather than spreading out, as quantum waves usually do. Taken at face value, this result suggested that the universe before the bounce was remarkably similar to our own: governed by general relativity and perhaps filled with stars and galaxies. If so, we should be able to extrapolate from our universe back in time, through the bounce, and deduce what came before, much as we can reconstruct the paths of two billiard balls before a collision based on their paths after the collision. We do not need to know each and every atomic-scale detail of the collision.

Unfortunately, various subsequent analyses dashed this hope. The model as well as the quantum waves used in the numerical simulations turned out to be a special case. In general, I found that waves spread out and that quantum effects

LILA RUBENSTEIN (balloons)

© 2014 Scientific American



Alternatively, before the big bounce the universe may have been in an almost unimaginable quantum state, devoid of space and time, when something triggered the big bounce and the formation of the atoms of spacetime. Which of these two alternatives occurred depends on further details that physicists are still working out.

were strong enough to be reckoned with. So the bounce was not a brief push by a repulsive force, like the collision of billiard balls. Instead it may have represented the emergence of our universe from an almost unfathomable quantum state—a world in highly fluctuating turmoil. Even if the preexisting universe was once very similar to ours, it passed through an extended period during which the density of matter and energy fluctuated strongly and randomly, scrambling everything—even the passage of time itself.

The fluctuations before and after the big bang were not strongly related to each other. The universe before the big bang could have been fluctuating very differently than it did afterward, and those details did not survive the bounce. The universe, in short, has a profound case of forgetfulness. It may have existed before the big bang, but quantum effects during the bounce wiped out almost all traces of this prehistory.

SOME SCRAPS OF MEMORY

THIS PICTURE OF THE BIG BANG is subtler than the classical view of the singularity. Whereas general relativity simply fails at the singularity, loop quantum gravity is able to handle the extreme conditions there. The big bang is no longer a physical beginning or a mathematical singularity, but it does put a practical limitation on our knowledge. Whatever survives cannot provide a complete view of what came before.

Frustrating as this may be, it might be a conceptual blessing. In physical systems as in daily life, disorder tends to increase. This principle, known as the second law of thermodynamics, is an argument against an eternal universe. If order has been decreasing for an infinite span of time, the universe should by now be so disorganized that structures we see in galaxies as well as on Earth would be all but impossible. The right amount of cosmic forgetfulness may come to the rescue by presenting the young, growing universe with a clean slate irrespective of all the mess that may have built up before.

According to traditional thermodynamics, there is no such thing as a truly clean slate; every system retains a memory of its past in the configuration of its atoms [see "The Cosmic Origins of Time's Arrow," by Sean M. Carroll; SCIENTIFIC AMERICAN, June 2008]. But by allowing the number of spacetime atoms to change, loop quantum gravity allows the universe more freedom to tidy up than classical physics would suggest.

All that is not to say that cosmologists have no hope of probing the quantum-gravitational period. Gravitational waves and neutrinos are especially promising tools because they barely interact with matter and therefore penetrated the primordial plasma with minimal loss. These messengers might well bring us news from a time near to, or even before, the big bang.

One way to look for gravitational waves is by studying their imprint on the cosmic microwave background radiation [see "Echoes from the Big Bang," by Robert R. Caldwell and Marc Kamionkowski; SCIENTIFIC AMERICAN, January 2001]. If quantum-gravitational repulsive gravity drove cosmic inflation, these observations might find some hint of it. Theorists must also determine whether this novel source of inflation could reproduce other cosmological measurements, especially of the early density distribution of matter seen in the cosmic microwave background.

At the same time, astronomers can look for the spacetime analogues of random Brownian motion. For instance, quantum fluctuations of spacetime could affect the propagation of light over long distances. According to loop gravity, a light wave cannot be continuous; it must fit on the lattice of space. The smaller the wavelength, the more the lattice distorts it. In a sense, the spacetime atoms buffet the wave. As a consequence, light of different wavelengths travels at different speeds. Although these differences are tiny, they may add up during a long trip. Distant sources such as gamma-ray bursts offer the best hope of seeing this effect [see "Window on the Extreme Universe," by William B. Atwood, Peter F. Michelson and Steven Ritz; SCIENTIFIC AMERICAN, December 2007].

In the case of material atoms, more than 25 centuries elapsed between the first speculative suggestions of atoms by ancient philosophers and Einstein's analysis of Brownian motion, which firmly established atoms as the subject of experimental science. The delay should not be as long for spacetime atoms. SA

Martin Bojowald is the leading researcher on the implications of loop quantum gravity for cosmology. He is a faculty member of the Institute for Gravitation and the Cosmos at Pennsylvania State University. Bojowald was awarded the First Award of the Gravity Research Foundation Essay Competition in 2003 and the Xanthopoulos Prize of the International Society for General Relativity and Gravitation in 2007. Outside physics, he enjoys reading classical literature and long-distance running in the Appalachian Mountains of central Pennsylvania.

MORE TO EXPLORE

Quantum Gravity. Carlo Rovelli. Cambridge University Press, 2004.
The Universe: A View from Classical and Quantum Gravity. Martin Bojowald. Wiley, 2012.
Back to the Beginning of Quantum Space-time. Martin Bojowald in *Physics Today*, Vol. 66, pages 35–41; March 2013.

scientificamerican.com/magazine/sa

© 2014 Scientific American

© 2014 Scientific American

COSMOLOGY

IS THE UNIVERSE LEAKING ENERGY?

Total energy must be conserved. Every student of physics learns this fundamental law. The trouble is, it does not apply to the universe as a whole

By Tamara M. Davis

IN BRIEF

As the universe expands and distant galaxies recede from us, their light gets redshifted, thus becoming less energetic. **This seeming violation** of the principle of conservation of energy is actually not in contradiction to accepted physical laws.

According to the author, the proper interpretation shows that the energy of individual photons is conserved. And phenomena taking place inside the galaxy generally conserve energy.

Photograph by Mark Hooper

© 2014 Scientific American

ENERGY CAN NEITHER BE CREATED NOR DESTROYED. THIS PRINCIPLE, CALLED conservation of energy, is one of our most cherished laws of physics. It governs every part of our lives: the heat it takes to warm up a cup of coffee; the chemical reactions that produce oxygen in the leaves of trees; the orbit of Earth around the sun; the food we need to keep our hearts beating. We cannot live without eating, cars do not run without fuel, and perpetual-motion machines are just a mirage. So when an experiment seems to violate the law of energy conservation, we are rightfully suspicious. What happens when our observations seem to contradict one of science's most deeply held notions: that energy is always conserved?

Skip for a moment outside our Earthly sphere and consider the wider universe. Almost all of our information about outer space comes in the form of light, and one of light's key features is that it gets redshifted—its electromagnetic waves get stretched—as it travels from distant galaxies through our ever expanding universe, in accordance with Albert Einstein's general theory of relativity. But the longer the wavelength, the lower the energy. Thus, inquisitive minds ask: When light is redshifted by the expansion of the universe, where does its energy go? Is it lost, in violation of the conservation principle?

Modern physics has shown that when we move far from the comfort of our everyday lives to explore the extremes of time and space, many of our basic assumptions start to crumble. We know from Einstein that simultaneity is an illusion that changes based on the observer's perspective and that notions of distance and duration are also relative. We now also suspect that the apparent continuity of time and space may be as illusory as the deceptively smooth appearance of matter. What is there in physics that we can rely on? Which of our deeply held principles is pulling the wool over our mind's eye and blinding us to the deeper truths? We physicists spend our days challenging what is known and striving to discover where our knowledge is inadequate or just plain wrong. And history is littered with the debris of discarded misconceptions. Is conservation of energy one of those misguided ideas?

It is not. On the scale of individual photons, energy is always conserved, even as light gets redshifted. Likewise, for phenomena that take place within our galaxy, violations are virtually impossible and our cherished law remains on a sound foundation. But on a cosmological scale, energy becomes a subtle concept indeed, and that is where things start to get interesting.

THE METAPHOR OF THE UNIVERSE AS AN EXPANDING RUBBER BALLOON SHOULD BE TAKEN WITH A GRAIN OF SALT.

SYMMETRY AND CONSERVATION

NOT ONLY HAS CONSERVATION of energy been empirically validated many times over, but scientists also have good theoretical reasons to believe it. Our confidence comes from German mathematician Emmy Noether, who put conservation of energy on a secure footing nearly 100 years ago, when she discovered that all conservation laws are based on symmetries of nature.

Usually you think of symmetry as something you see in a mirror, a reflection of some kind or a rotation perhaps. An equilateral triangle is symmetric because you can flip it sideways or rotate it one third of the way around, and you end up with exactly the same shape. A square also has symmetry, but you need rotate it only one fourth of the way around to find an identical configuration. The most symmetric of the two-dimensional objects is the circle because you can rotate it any amount and reflect it over any axis through its center, and it remains exactly the same—it displays what is called continuous symmetry.

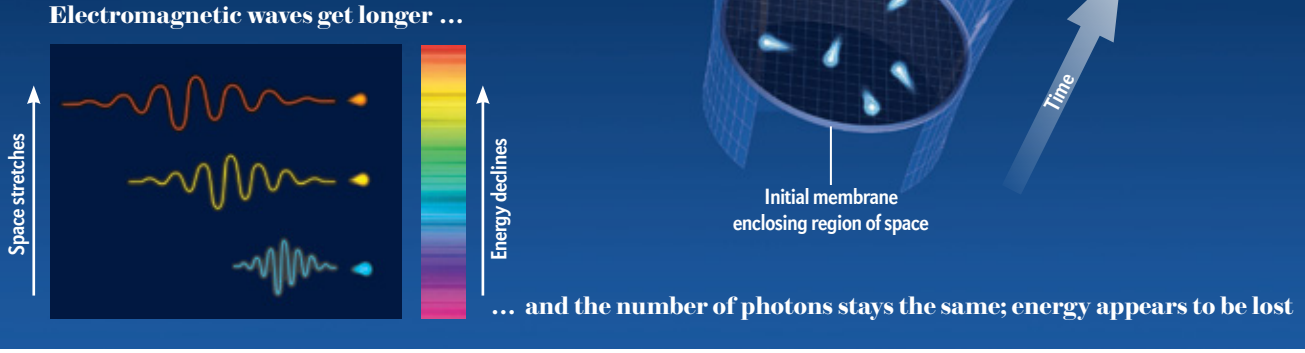
Physical laws, too, can be symmetric. The passage of time does not change the laws of nature; if you repeat an experiment many times—for example, making billiard balls collide at a given angle—the result is always the same. This quality is known as time symmetry. The laws of nature do not change depending on where you are—so we have spatial symmetry. Nor do the laws of nature change depending on the direction in which you look (rotational symmetry). Sure, the scenery may change depending on where you are standing, when you are standing

© 2014 Scientific American

Why Energy Seems to Disappear

Those who argue that the universe is losing energy base their conclusion in part on the redshift of light. The universe appears to be expanding, as if space itself were getting stretched out. In consequence, the electromagnetic waves that compose light get stretched as well, shifting, in the case of visible light, toward the red part of the spectrum (*below*). Photons of longer wavelength have lower energy, so logic dictates that each photon must become less energetic as it travels toward us.

But does the universe as a whole lose energy? The total energy of the photons in the universe cannot be calculated, but one can in principle calculate the energy contained within an imaginary membrane that expands in concert with the universe. (At the right, the region inside a membrane is represented as two-dimensional.) Photons can enter or exit through the membrane, but the uniform density of space tells us that the number of photons in the enclosed region will roughly stay constant. Because each photon in the region becomes less energetic as space expands, this calculation suggests that the total amount of photon energy in the region and, by implication, in the rest of the universe must be going down.


Electromagnetic waves get longer …
Space stretches
Energy declines
Photons
Expanded membrane
Time
Initial membrane enclosing region of space
… and the number of photons stays the same; energy appears to be lost

there and the direction you are looking, but the fundamental underlying laws of physics that dictate how that scenery behaves are independent of your location, orientation and time. When a law remains unchanged regardless of the situation, it, like the circle, is said to be continuously symmetric.

What Noether discovered is that whenever nature displays a continuous symmetry, a conservation law comes along for the ride, and vice versa. In particular, spatial symmetry dictates that momentum is conserved; rotational symmetry ensures angular momentum is conserved; and time symmetry means that energy is conserved.

So, saying that energy is conserved is as solid as saying that the laws of physics are the same now as they were in the past and will be in the future. On the other hand, were time symmetry to break down, conservation of energy would fail. As we will see, this is where energy conservation may start to get in trouble in Einstein's universe.

GO WITH THE FLOW

THERE IS NO BETTER WAY to test whether the present matches the past, and thus to see if energy is conserved in the universe, than to watch the past in full live action through an astronomer's telescope. Our telescopes are now so powerful that we are able to see back to when the first galaxies were forming and beyond to the piping-hot afterglow of the big bang itself. The light we are seeing has been traveling for billions of years, and in all that time, the first thing it has hit is the mirror of our telescope. The wavelengths of that light are our key to assessing conservation.

In the 1920s Edwin Hubble discovered that the light of most galaxies is redshifted: he found that the wavelengths of photons

Illustrations by Moonrunner Design

© 2014 Scientific American

that were emitted or absorbed by atoms (such as by hydrogen) in all but the nearest galaxies to us appear, when they reach us, to be stretched relative to the wavelengths emitted by the same atoms at home—stretching roughly in proportion with the galaxies' distance. In fact, ever since the discovery of this phenomenon, whenever astronomers cannot measure a galaxy's distance more directly they instead give an estimate using its redshift as a proxy.

Redshifts (and blueshifts) also happen all the time here on Earth. Imagine driving past a police radar. As your car approached, the electromagnetic waves from the radar would look slightly shrunk to you—if you could see them—when they reached you. But after you passed, the waves would look a bit stretched. This is the Doppler effect: it is the electromagnetic equivalent of the familiar change in acoustical pitch you would hear in a siren as it passes by. (The police officer can tell if you are speeding by measuring a Doppler shift in the reflection of the radar.) Although in this case the waves are not in the visible spectrum, physicists still call the stretching and shrinking of the waves redshift and blueshift, respectively.

Cosmological redshifts, however, are generally considered to be different from the Doppler effect. Doppler shifts are caused by relative motion. In that case, the photons are not losing or gaining energy; they just look different to you than they do to the emitter. In contrast, most general relativity or cosmology textbooks say cosmological redshifts happen because as light travels, the very space it travels in gets stretched like the surface of an inflating rubber balloon.

In fact, cosmological redshifts can happen even when there seems to be no relative motion at all, as the following thought experiment shows. Imagine a galaxy far, far away but connected to ours by a long tether. Relative to us, the galaxy is not moving, even as other galaxies in its vicinity recede from us. Yet standard calculations show that the light reaching us from the tethered galaxy will still be redshifted (though not quite as strongly as the light from the galaxies in its vicinity, which have not been pulled out of the flow of the expansion). This redshift is usually attributed to the stretching of the space through which light travels.

More Cosmic Puzzles

Is space within our galaxy expanding? No. Cosmic-scale expansion does not affect the dynamics inside a galaxy. Once local gravitational effects cause a galaxy to form, the expansion has no power to pull the galaxy apart.

Do photons from distant galaxies get redshifted because the universe's density has been decreasing? After all, photons get redshifted when they climb up a gravitational gradient. True, but at any given time, the universe was uniform, so the density of matter was the same behind a photon as it was in front of it. Thus, photons had no gravitational gradient to climb out of.

Is entropy compatible with time symmetry? Yes. In complex interactions of particles, such as the breaking of an egg, we can tell which way a movie of the process is being played—the direction in which entropy increases, which is the direction of increasing disorder. Nevertheless, any single one of the interactions between particles could happen forward or backward, as far as the laws of physics are concerned.

PECULIAR PHYSICS

SO PHOTONS TRAVELING in an expanding universe appear to lose energy. What about matter? Does it lose energy, too? When we describe the motion of matter in the universe, we distinguish between two different types. An object can just be receding with the general flow of the universe's expansion, just like dots painted on our balloon would recede from one another as the balloon inflates. In cosmology, such an object is called comoving. But an object can also have its own motion on top of the motion caused by the expansion. This second type is called peculiar motion, and it takes place when something is dragged out of the smooth flow of the expansion by local effects, such as the gravitational pull of a nearby galaxy or the thrust of a rocket.

Galaxies themselves always have at least a bit of peculiar motion, but for distant galaxies—which recede faster than near ones do—the peculiar velocity is small compared with their recession. At the largest scales, the distribution of galaxies is uniform, so local effects are negligible and galaxies are essentially comoving. They can be regarded as the dots on the balloon, that is, as flag posts of the expanding fabric of space.

A comoving frame of reference such as that defined by galaxies is very handy: for example, it gives a universal convention for time, so that everyone in every comoving galaxy would agree on how long ago the big bang happened.

If an intergalactic traveler drifts for billions of light-years, he or she will pass many of these flag-post galaxies. But because the universe is expanding, the flag posts are moving away from one another, and our traveler appears to be going slower and slower relative to each subsequent galaxy he or she passes. So the traveler appears to slow down.

Thus, much as light loses energy by increasing in wavelength, matter loses energy by slowing down. At first sight those behaviors appear to be very different. But interestingly, quantum mechanics unifies the two. In the quantum-mechanical view of matter, particles that have mass also have wavelike properties. French physicist Louis de Broglie found that the larger the momentum of a particle, the smaller its wavelength and the greater its energy—and he won the Nobel Prize in 1929 for his discovery.

Particles of matter can have high momentum by having high mass or high velocity, or both. That feature explains, for example, why a baseball does not appear to wiggle about in wavelike motions after it leaves the pitcher's glove. Baseballs are enormously massive in quantum terms, and at the typical speed of a major-league fastball pitch (about 145 kilometers an hour) a baseball has a wavelength of 10^{-34} meter—not something a batter will have to worry about. On the other hand, an electron traveling at the same speed has a wavelength of 18 microns: still small, but 29 orders of magnitude larger than a baseball's, and very noticeable when it comes to the behavior of electrons.

When you calculate how much relative velocity massive particles lose as they pass by their receding neighbors, you find that the de Broglie wavelength of the particles increases by exactly the same proportion as a photon's wavelength does.

© 2014 Scientific American

A Matter of Changing Geometry

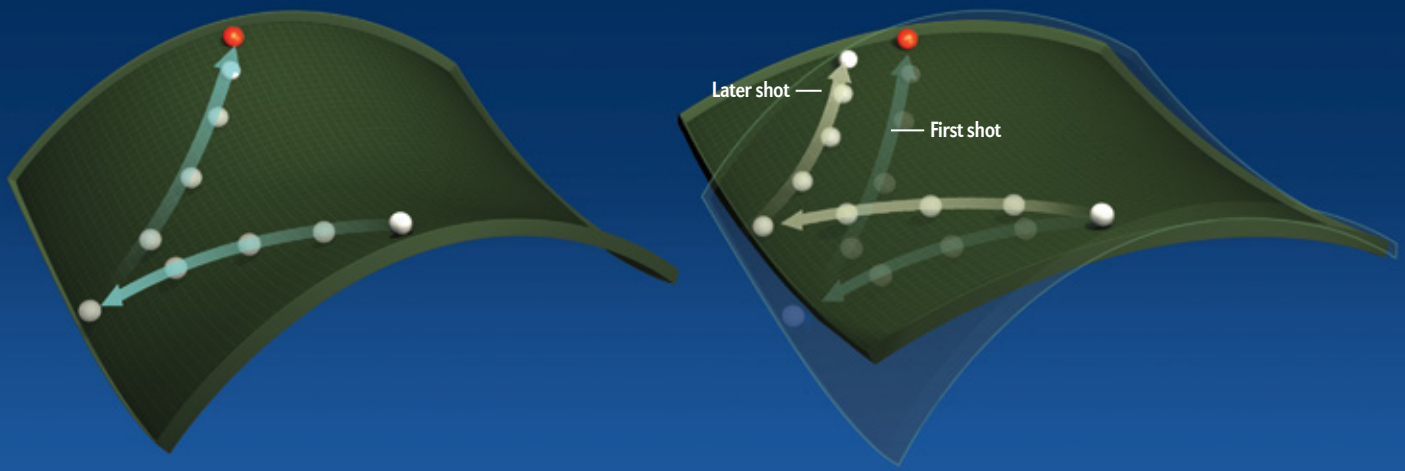
Conservation laws are intimately tied to symmetries of nature. In particular, energy is conserved when the laws of nature have "time symmetry." Time symmetry is said to occur if any experiment gives the same results regardless of when it is performed. But if experiments can give different results at different times, energy may not be conserved. An example is playing a bank shot on a pool table that has a changing geometry. At cosmological scales, our universe has a changing geometry, which once again implies that energy may not be conserved.

Curved Pool Table

To play on a table that has curved, or "non-Euclidean," geometry, you have to adjust your shots to the geometry. Still, if the geometry is fixed, the same exact shot will work again in the future. Because of this time symmetry, in a universe with fixed geometry energy would be conserved.

Evolving Geometry

If the pool table has a geometry that changes in time, however, the shots that worked in the past may not work again—therefore, time symmetry is broken. Something similar can happen in the universe because according to general relativity the motion of matter and energy changes the geometry of space. Under these conditions, energy need not be conserved.



Thus, light and matter seem to behave in exactly the same way when it comes to energy loss in the expanding universe, and in both cases it looks as if energy conservation is being violated. In the case of matter, the paradox is explained by the fact that we are measuring velocity in different frames of reference—that is, relative to the receding galaxies. As we will see, something similar happens with photons.

CREATIVE ACCOUNTING

WERE COSMOLOGICAL ACCOUNTANTS to verify that the universe is losing energy, they might attempt to tally up all the energy in the universe, rather than focusing on one object at a time. They might first add up all the energy contained simply in the mass of the matter in the universe (mass m and energy E are equivalent following Einstein's $E = mc^2$, where c represents the speed of light). Then they would add in the kinetic energy related to the matter's peculiar motion. To that sum, they would have to add the energy of light as well and then get to the complex job of counting the energy in all the gravitational fields around planets, stars and galaxies, plus the energy contained in chemical bonds and in the nuclei of atoms. (Sound and heat are just the motion of particles, so they have already been accounted for.)

A first problem they would face is that the universe may be infinitely large and contain an infinite amount of matter and energy. Thus, the accountants would need to take a shortcut. They would draw an imaginary membrane around a region of the universe and add up the energy inside [*see box on page 91*]. They then would let the membrane expand as the universe does, so that comoving galaxies stay inside the membrane. Light and matter can pass in and out of the membrane, but because the universe is homogeneous, the same amount leaves as enters, so the amount inside the membrane stays roughly constant. Our accountants know that the whole universe can be constructed from a series of such volumes. If the energy in the universe is to be conserved as a whole, therefore, it is enough to show that the energy in any one of those volumes is conserved.

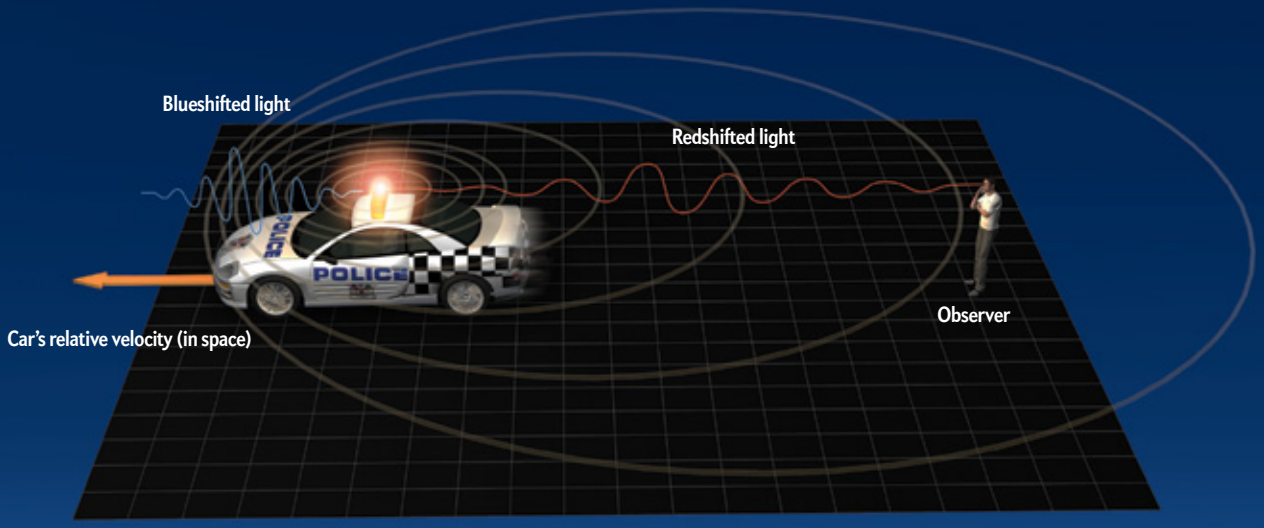
The calculation is easy to do for matter that is at rest—just chilling out and going with the flow of the expansion. Its only energy in this case comes from its mass, and because no matter leaves or enters the membrane, we know the mass is conserved. But things are a bit more complicated for light, as we have seen, and for matter that has peculiar velocity. Although the number of photons or of matter particles within the membrane does not change, over time photon energy is lowered, as is the kinetic energy of the peculiarly moving matter. Therefore, the total energy in the membrane goes down.

The situation would be even more complicated if the accountants were to count dark energy, which is what is causing the universe's expansion to accelerate. The nature and properties of dark energy are still a mystery, but it appears that dark energy

© 2014 Scientific American

How Photon Energy Is Conserved After All

The redshift we see in distant galaxies is usually attributed to the stretching of space, but it can also be interpreted as an effect of the receding motion of the galaxies with respect to the observer. It is therefore similar to the familiar Doppler effect, which one can hear in the siren of a police car that is passing by but that also affects the wavelengths of photons—for example, those from the car's emergency lights (*below*). In the case of the police car, energy is conserved; similarly, calculating galaxy redshift as a Doppler shift (*opposite page*) shows that photons from a distant galaxy also do not lose energy.



Ordinary Doppler Shift

Doppler shifts arise from relative motion. The lights flashing from the top of a police cruiser appear redshifted or blueshifted—though imperceptibly to human eyes—depending on whether the car is moving away from you or toward you. The larger the car's velocity relative to an observer, the stronger the effect will be. But the occurrence of the Doppler shift does not mean that photons change color (nor that they lose energy) along the way; they just have different colors as seen from an observer's point of view than they have from the car's own point of view.

does not dilute as the universe expands. Thus, as the volume in our membrane increases, the amount of energy in that volume increases as well, with the additional energy seemingly coming out of nowhere! One might think that the increase in dark energy could balance out the losses in all other forms of energy, but that is not the case. Even if we take dark energy into account, the total energy within the membrane is not conserved.

How do our accountants reconcile these changing energies with Noether's theorem? In fact, they would soon realize that there is no reason why Noether's theorem should apply to our changing universe. According to general relativity, matter and energy curve space, and as matter and energy move (or spread out in an expanding space) the shape of space changes accordingly. In everyday life, these effects are essentially too small to detect, but at cosmic scales they can be relevant.

This malleability of space implies that the behavior of the universe is not time-symmetric. The easiest way to visualize this fact is to go back to the example of the billiard balls. If we watched several movies of a particular shot being played on a pool table of changing geometry—for example, one that starts flat and warps with time—each movie would look different from the others; you could tell when and in what order each movie was taken. Time symmetry would be broken [*see box on preceding page*].

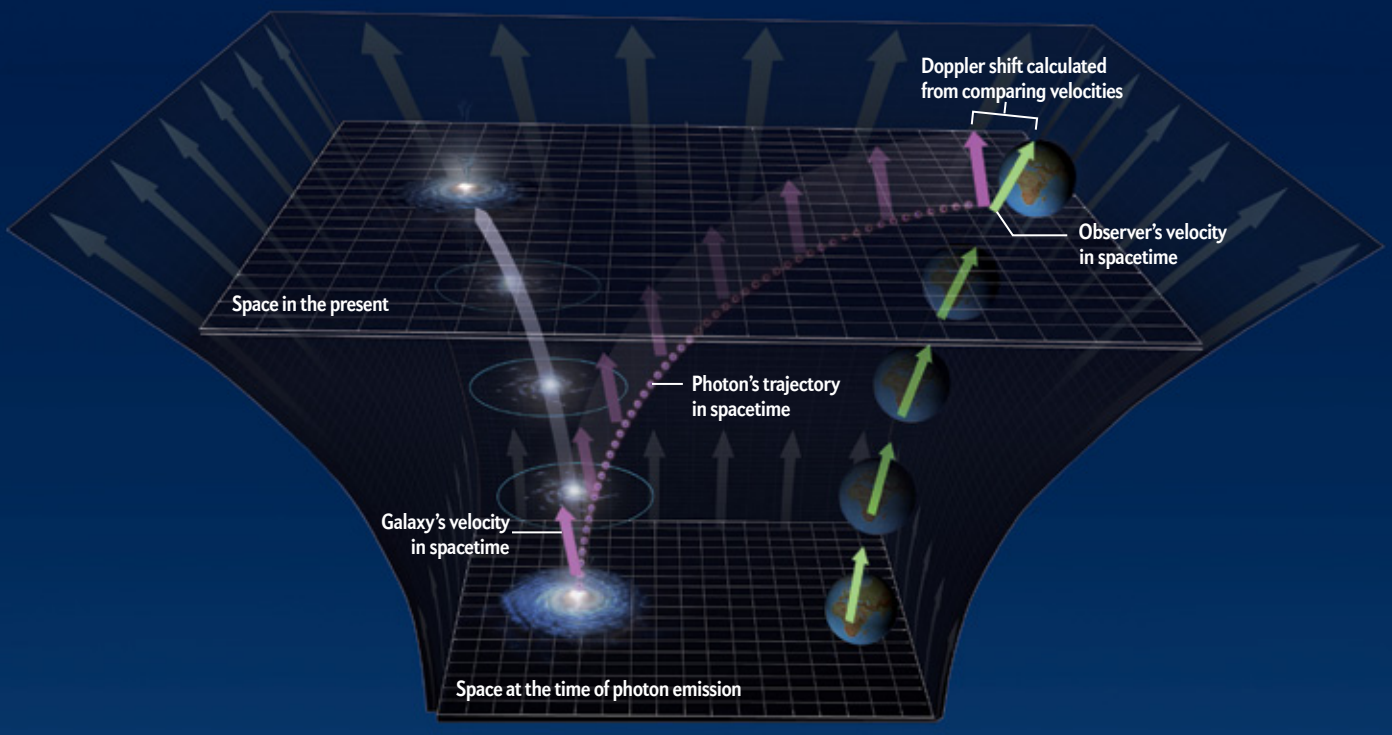
We have come to the limit of our cherished conservation principles: when time and space themselves are mutable, time symmetry is lost, and conservation of energy need no longer hold.

COSMIC SEMANTICS

EVEN IF CURVATURE does not change, however, trying to tally up the energy of the universe is a futile exercise: our accountants' godlike perspective does not pertain to any observer in the universe. In particular, they do not take into account the energy of comoving galaxies' motion with respect to one another, so to them, the galaxies appear to have no kinetic energy. Another issue is the gravitational energy associated with the galaxies' mutual attraction. A well-known problem with general relativity is that in the theory one cannot always unambiguously define gravitational energy in a way that applies to the universe as a whole.

Thus, the total energy of the universe is neither conserved nor lost—it is just undefinable. On the other hand, if we abandon the godlike point of view and instead focus on one particle at a time, we can find what many cosmologists believe is a more natural way of thinking of the journey of a photon from a distant galaxy. In this interpretation, the photon does not lose energy after all. The point is that our metaphor of the expanding rubber balloon, though useful to visualize the expansion, should be taken with a grain of salt: empty space does not have a physi-

© 2014 Scientific American


Doppler shift calculated from comparing velocities
Observer's velocity in spacetime
Space in the present
Photon's trajectory in spacetime
Galaxy's velocity in spacetime
Space at the time of photon emission

Galaxy Redshift as a Doppler Shift

A galaxy's redshift is identical to the Doppler shift an observer would see when watching a police car recede at the same relative velocity as the galaxy—as long as "relative velocity" is interpreted in the appropriate way. First, one must trace the trajectories of the galaxy and of the observer not in space but in spacetime. (In the schematic view here, space is an evolving two-dimensional surface; spacetime trajectories cut through it.) Second, one must compare the velocity of the galaxy at the time when it emitted the photon (*purple arrow*) with the velocity of the observer at the time when the photon was received (*green arrow*) and then—using the appropriate math derived from general relativity—calculate the relative velocity. The Doppler shift calculated from this relative velocity coincides with the galaxy's redshift, suggesting that the galaxy's redshift can be interpreted as the result of relative motion, rather than of the expansion of space. Therefore, no energy is lost.

cal reality. As galaxies recede from one another, we are free to consider this relative motion as "expansion of space" or as "movement through space"; the difference is mostly semantics.

Cosmological redshift is usually described as a consequence of the expansion of space. But in Einstein's general relativity, space is relative, and what really matters is a galaxy's history—the trajectory it describes in space*time*. Thus, we should calculate the relative velocity of the distant galaxy with respect to us by comparing its trajectory in spacetime with ours. The amount of redshift seen in the galaxy turns out to be identical to the Doppler shift the observer would see in a car that is receding at the same relative velocity [*see box above*].

This happens because in small enough regions the universe makes a pretty good approximation of flat spacetime. But in flat spacetime there is no gravity and no stretching of waves, and any redshift must just be a Doppler effect. So we can think of the light as making many tiny little Doppler shifts along its trajectory. And just as in the case of the police car—where it would not even occur to us to think that photons are gaining or losing energy—here, too, the relative motion of the emitter and observer means that they see photons from different perspectives and not that the photons have lost energy along the way.

In the end, therefore, there is no mystery to the energy loss of photons: the energies are being measured by galaxies that are receding from one another, and the drop in energy is just a matter of perspective and relative motion.

Still, when we tried to understand whether the universe as a whole conserves energy we faced a fundamental limitation because there is no unique value we can ever attribute to something called the energy of the universe.

Thus, the universe does not violate the conservation of energy; rather it lies outside that law's jurisdiction. SA

Tamara M. Davis is an associate professor at the University of Queensland in Brisbane, Australia. With the universe as her laboratory, she works with large observational cosmology teams that use supernovae and the distribution and motion of galaxies to learn about the nature of gravity, spacetime and our fundamental laws of physics. You will also often find her playing ultimate Frisbee—she has represented Australia at four World Championships, twice as team captain.

MORE TO EXPLORE

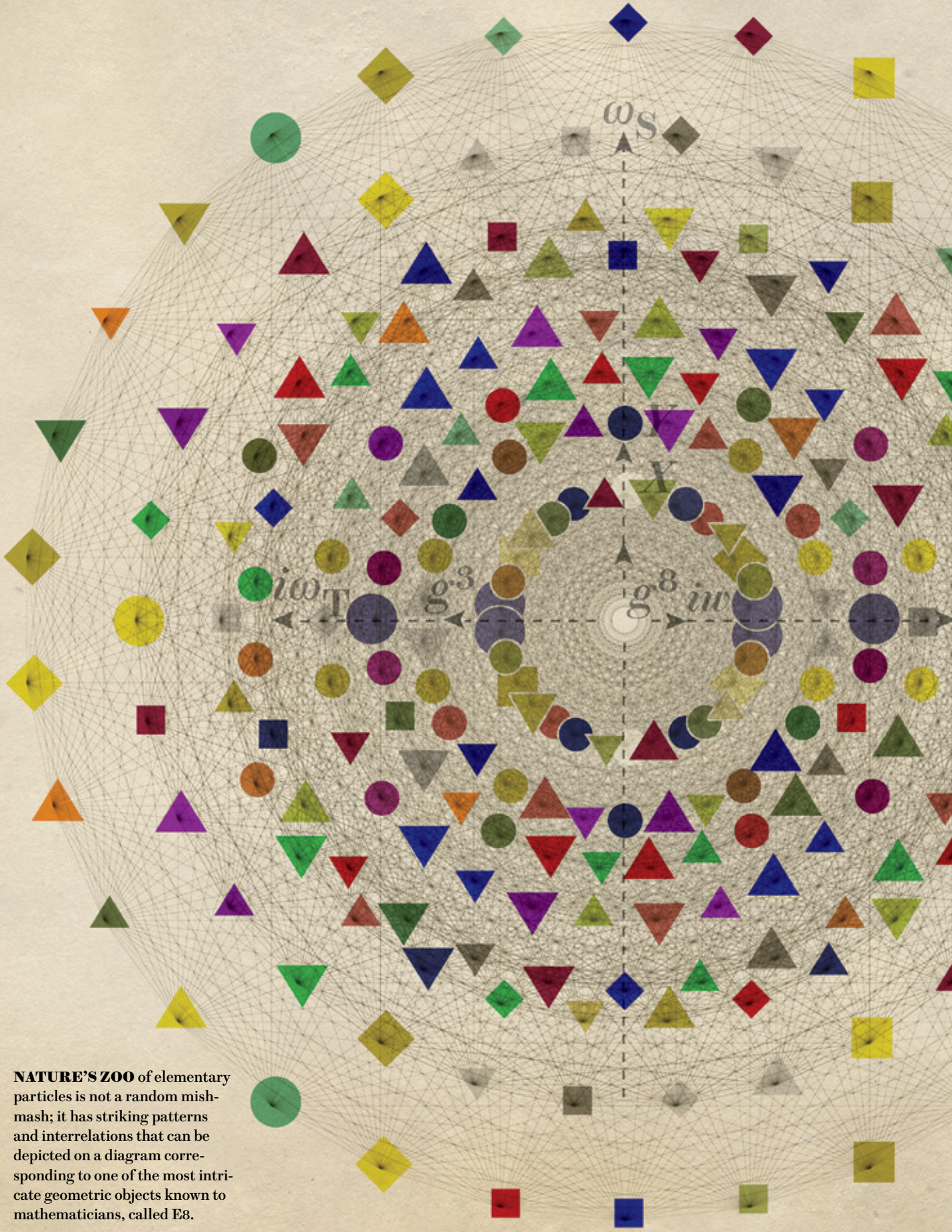
Spacetime and Geometry: An Introduction to General Relativity. Sean M. Carroll. Addison-Wesley, 2003.

Misconceptions about the Big Bang. Charles H. Lineweaver and Tamara M. Davis in *Scientific American*, Vol. 292, No. 3, pages 36–45; March 2005.

The Kinematic Origin of the Cosmological Redshift. Emory F. Bunn and David W. Hogg in *American Journal of Physics*, Vol. 77, No. 8, pages 688–694; August 2009.

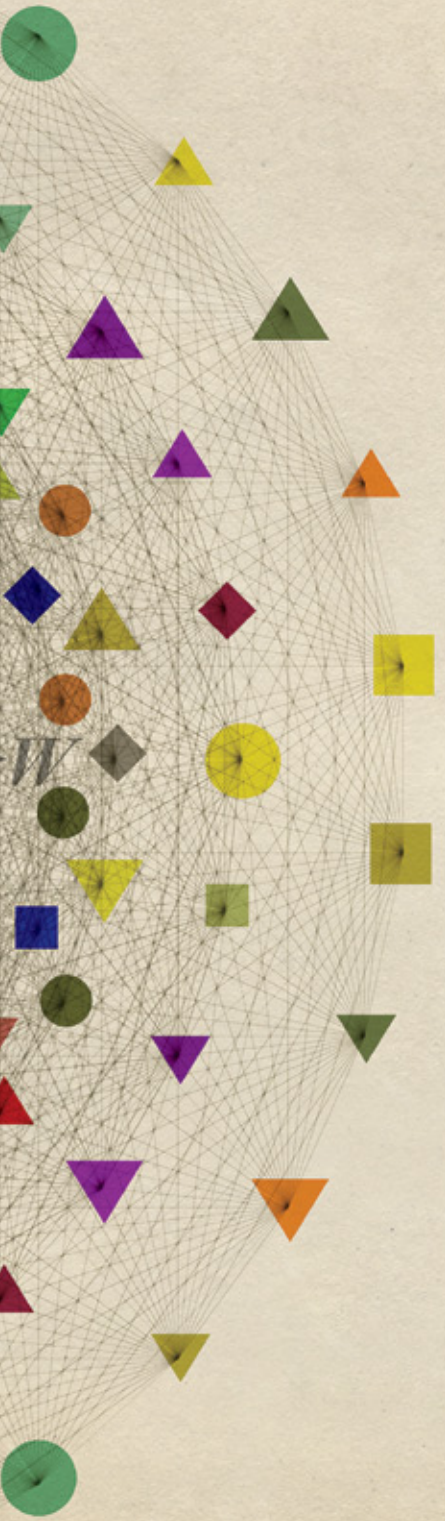
scientificamerican.com/magazine/sa

© 2014 Scientific American



NATURE'S ZOO of elementary particles is not a random mish-mash; it has striking patterns and interrelations that can be depicted on a diagram corresponding to one of the most intricate geometric objects known to mathematicians, called E8.

© 2014 Scientific American



GRAND UNIFIED THEORY

A GEOMETRIC THEORY OF EVERYTHING

Deep down, the particles and forces of the universe are a manifestation of exquisite geometry

By A. Garrett Lisi and James Owen Weatherall

MODERN PHYSICS BEGAN WITH A SWEEPING UNIFICATION: IN 1687 ISAAC Newton showed that the disparate theories describing everything from planetary motion to tides to pendulums were all aspects of a universal law of gravitation. Unification has played a central role in physics ever since. In the middle of the 19th century James Clerk Maxwell found that electricity and magnetism were two facets of electromagnetism. One hundred years later electromagnetism was unified with the weak nuclear force governing radioactivity, in what physicists call the electroweak theory.

This quest for unification is driven by practical, philosophical and aesthetic considerations. When successful, merging theories leads us to discover things we might otherwise never have suspected. Much of the activity in experimental particle physics today, at accelerators such as the Large Hadron Collider at CERN near Geneva, involves a

IN BRIEF

In 2007 physicist A. Garrett Lisi wrote the most talked about theoretical physics paper of the year. Outlets from the *New Yorker* to *Outside* magazine were drawn to the story, partly on account of his surfer lifestyle. Lisi and others have continued to develop the theory.

Most physicists think reconciling Einstein's general theory of relativity with quantum theory will require a radical shift in our conception of reality. Lisi, in contrast, argues that the geometric framework of modern quantum physics can be extended to incorporate Einstein's theory, leading to a long-sought unification of physics.

Even if Lisi turns out to be wrong, the E8 theory he has pioneered showcases striking patterns in particle physics that any unified theory will need to explain.

© 2014 Scientific American

Illustrations by Chad Hagen

search for novel phenomena predicted by the unified electroweak theory. In addition to predicting new physical effects, a unified theory provides a more elegant picture of how our universe operates. Many physicists believe all physical phenomena match the patterns of some beautiful mathematical structure.

The current best theory of nongravitational forces—the electromagnetic, weak and strong nuclear force—was largely completed by the 1970s and is known as the Standard Model. The theory describes these forces and particles as the dynamics of geometric objects called Lie groups and fiber bundles. It is, however, a patchwork; a separate geometric object governs each force. Physicists have proposed various Grand Unified Theories, or GUTs, in which one geometric object would explain all these forces, but no one knows which of these GUTs is true.

And an even deeper unification problem faces today's physicists. In a fully unified theory, gravity and matter should also combine naturally with the other forces, all as parts of one mathematical structure—a Theory of Everything. Since the 1980s string theory, the dominant research program in theoretical particle physics, has been an attempt to describe gravity and the Standard Model using elaborate constructs of strings and membranes vibrating in many spacetime dimensions.

But string theory is not the only effort. An alternative, loop quantum gravity, uses a more minimal framework, closer to that of the Standard Model [see "Atoms of Space and Time," by Lee Smolin; Scientific American, January 2004]. Building on its insights, one of us (Lisi) proposed a new unified theory in 2007. The basic idea is to extend Grand Unified Theories and include gravity as part of a consistent geometric framework. In this unified field theory, called E8 theory, all forces and matter are described as the twisting of a single geometric object.

Many physicists are skeptical—and rightly so. The theory remains incomplete. But even in this early stage of development, it unveils some of the beautiful structures in play at the deepest levels of nature, and it makes predictions for new particles that the Large Hadron Collider might find. Although physicists are not yet at the culmination of our centuries-long quest for unity, E8 theory is an important step on that journey.

EVERY FIBER OF OUR BEING

TO DESCRIBE E8 THEORY, we first must set out the geometric principles that govern all known forces and particles. Geometry is the study of shape, but in the case of physics, you might wonder: shape of what? Plato thought elements such as earth and air were associated with little cubes and octahedra. Similarly, today the geometric objects associated with elementary particles are perfect, smooth shapes, existing outside our space yet connected to it. We cannot see them directly, but we see their effects.

The main geometric idea underlying the Standard Model is that every point in our spacetime has shapes attached to it, called fibers, each corresponding to a different kind of particle. You can envision the universe as a Chia Pet (a terra-cotta figurine covered with sprouts). The surface of the figurine is analogous to spacetime, and the sprouts are fibers. The entire geometric object—spacetime and fibers together—is called a fiber bundle. The fibers are not in our space but over it; they may be thought of as different, internal spaces attached to each point of our spacetime, with shapes corresponding to particles' properties.

This idea, introduced in 1918 by mathematician Hermann Weyl, is now a well-established principle of physics [see "Fiber Bundles and Quantum Theory," by Herbert J. Bernstein and Anthony V. Phillips; Scientific American, July 1981]. Distinct from the speculated undulating spatial dimensions of string theory, these internal space fibers are of fixed shape. Their dynamics arise from how they are attached to four-dimensional spacetime.

The electric and magnetic fields existing everywhere in our space are the result of fibers with the simplest shape: the circle. A circle, called U(1) by physicists, is the simplest example of a Lie group (pronounced "Lee," after 19th-century Norwegian mathematician Sophus Lie). It has a single symmetry: if we rotate a circle, it remains the same. A small rotation like this is called a generator of the Lie group. Following a generator, just like drawing with a compass, takes us around a circle.

The fiber bundle of electromagnetism consists of circles attached to every point of spacetime [*see box on opposite page*]. Crucially, each circle can rotate a little relative to its spacetime neighbors. The so-called connection field of a fiber bundle describes how neighboring fibers are related by these symmetry rotations. The electric and magnetic force fields filling spacetime correspond to the curvature of this fiber bundle—geometrically, the electric and magnetic fields are how the circular fibers twist over time and space. An electromagnetic wave is the undulation of circles over spacetime. One quantum of an electromagnetic wave—a photon—is a propagating particle of light.

Each kind of elementary particle corresponds to a different fiber over spacetime; the Chia Pet has many different kinds of sprouts. All electrons result from the twisting of a single kind of fiber—explaining, among other things, why all electrons are identical. The fibers of electrically charged particles, such as electrons, wrap around the circular fibers of electromagnetism like threads around a screw. How fast a particle's fiber twists around the circle is equal to its electric charge, determining how the particle responds to the force of electromagnetism.

Because twists must meet around the circle, these charges are integer multiples of some standard unit of electric charge. Of the elementary matter particles, called fermions, electrons have electric charge –1 (three twists), up quarks have electric charge +⅔ (two opposite twists), down quarks have electric charge –⅓ (one twist), and neutrinos have 0. Antimatter particles, such as positrons, have twists in the opposite direction around the electromagnetic circle, giving them the opposite electric charges.

When particles collide, they may be converted into new types, but the outgoing particles have exactly the same total charge as the incoming ones did. This crucial fact is a consequence of fiber geometry: When any two particles meet, their twists add. This fiber-bundle picture explains much of electromagnetism. The electric charges describe the geometric structure of the combined electromagnetic and matter fiber bundle, determining the possible interactions between electrically charged particles.

DIFFERENT CHARGES FOR DIFFERENT FORCES

PHYSICISTS APPLY these same principles to the weak and strong nuclear forces. Each of these forces has its own kind of charge and its own propagating particles. They are described by more complicated fibers, made up of sets of intersecting circles, interacting with themselves and with matter according to their twists.

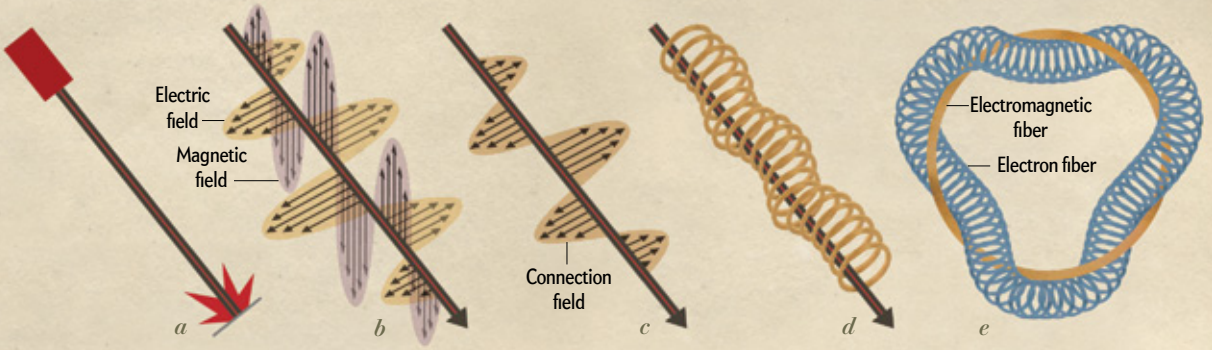
The weak force is associated with a three-dimensional Lie group fiber called SU(2). Its shape has three symmetry genera-

© 2014 Scientific American

BASICS

From Electromagnetism to Geometry

The geometric view of nature follows naturally from the way the world around us works. The simplest and most familiar examples are the forces of electricity and magnetism. Electric sparks, magnetic attraction and laser light are different manifestations of the electric and magnetic fields that pervade space. In fact, physicists think that everything in the world—all the forces of nature and even all the particles of matter—arises from different kinds of fields. The behavior of these fields hints at an underlying geometric structure.



A laser beam (*a*) consists of oscillating electric and magnetic fields (*b*), which are the manifestations of a single electromagnetic "connection" field (*c*). This field describes how circular "fibers" of electromagnetism twist across the beam. That is, a circular fiber is attached to every point in spacetime, and a particle of light (a photon) corresponds to the undulation of these circles (*d*). A charged particle such as an electron corresponds to another fiber that wraps around the circular fiber (*e*).

tors, corresponding to the three weak-force boson particles: W^+, W^- and W^3—relatives of the photon. Each Lie group is a multidimensional, smooth tangle of intersecting circles twisting around one another. The circles of the W^+ and W^- bosons in SU(2) twist oppositely around the W^3 circles and so have weak charge, W, of +1 and –1. Because they have weak charge, these particles interact with one another as well as with matter.

Exactly half of elementary matter particles interact with the weak force, their fibers twisting around the W^3 and other circles of SU(2). Fermions come in two varieties, related to how their spin aligns with their momentum: left-handed and right-handed. Only the left-handed fermions have weak charges, with the left-handed up quark and neutrino having weak charge +½ and the left-handed down quark and electron having weak charge –½. For antiparticles, this is reversed: only right-handed antiparticles have weak charge. In other words, our universe is not left-right symmetrical—we can tell whether we are looking at weak interactions directly or looking at them in a mirror. This asymmetry is one of many mysteries a unified theory seeks to explain.

When physicists unified the weak force with electromagnetism to create the electroweak theory, they combined the SU(2) fiber with a U(1) circle. This circle is not the same as the electromagnetic one; it represents a precursor to electromagnetism known as the hypercharge force, with particles twisting around it according to their hypercharge, labeled Y. Inside the combined four-dimensional electroweak Lie group, the W^3 circles combine with the hypercharge circles to form a two-dimensional torus. This torus can be sliced in many ways, just as every person has their own idiosyncratic way to slice a bagel. The fibers of particles known as Higgs bosons twist around the electroweak Lie group and determine a particular set of circles, breaking the symmetry—like someone insisting there is only one true way to cut a bagel. The Higgs does not twist around these circles, which then correspond to the massless photon of electromagnetism.

Perpendicular to these circles are another set that should correspond to another particle, which the developers of electroweak theory called the Z boson. The fibers of the Higgs bosons twist around the circles of the Z boson, as well as the circles of the W^+ and W^-, making all three particles massive. Experimental physicists discovered the Z in 1973, vindicating the theory and demonstrating how geometric principles have real-world consequences.

To see how the electroweak theory works, we can plot the weak charges and hypercharges of all known particles [*see box on next four pages*]. Because mathematicians call charge "weight," this plot is known as a weight diagram. In it, all particles line up on equally spaced oblique lines, corresponding to their electric charges. Electric charge is thus a specific combination of weak charge and hypercharge, determined by the Higgs bosons. By measuring the strength of the weak force, physicists know that the angle of these lines, known as the weak mixing angle, is about 30 degrees. Explaining the value of this angle is one of the most tangible goals of a unified theory of physics.

COLORFUL PHYSICS

IN THE STANDARD MODEL, the strong nuclear force that binds quarks into atomic nuclei corresponds geometrically to an even larger Lie group, SU(3). The SU(3) fiber is an eight-dimensional internal space composed of eight sets of circles twisting around one another in an intricate pattern, producing interactions among eight kinds of photonlike particles called gluons on account of how they "glue" nuclei together. As complicated as this fiber shape is, we can break it into comprehensible pieces. Embedded within it is a torus formed by two sets of untwisted circles, corresponding to two generators, g^3 and g^8. The remaining six gluon generators twist around this torus, and their resulting g^3 and g^8 charges form a hexagon in the weight diagram.

© 2014 Scientific American

The quark fibers twist around this SU(3) Lie group, their strong charges forming a triangle in the weight diagram. These quarks are whimsically labeled with three colors: red, green and blue. A collection of matter fibers forming a complete pattern, such as three quarks in a triangle, is called a representation of the Lie group. The colorful description of the strong interactions is known as the theory of quantum chromodynamics.

Together quantum chromodynamics and the electroweak model make up the Standard Model of particle physics, with a Lie group formed by combining SU(3), SU(2) and U(1), as well as matter in various representations. This structure is described by a weight diagram with four charge axes, which may be projected down to two dimensions and plotted. This diagram displays the crown jewels of modern physics. Every allowed particle interaction of the Standard Model may be found on it.

The Standard Model is a great success. But it presents several puzzles: Why does nature use this combination of Lie groups? Why do these matter fibers exist? Why do the Higgs bosons exist? How is gravity included? And there are other mysteries. The quarks, electrons and neutrinos that constitute common matter are called the first generation of fermions; they have second- and third-generation doppelgängers with identical charges but much larger masses. Why is that? And what are cosmic dark matter and dark energy? A unified theory should be able to provide answers to these and other questions. The first step toward such a theory is the unification of electroweak and strong forces.

GRAND (BUT NOT FULL) UNIFICATION

ALTHOUGH THE ELECTROWEAK and strong forces can both be described using fiber bundles, their fibers are separate. Physicists have asked whether some single fiber encompasses both. Instead of different Lie groups for each force, there would be a single, larger Lie group for all. They have good evidence for this idea: all these forces become close in strength at very short distances, indicating they are aspects of a single force. A Grand Unified Theory would describe this force, reproduce the Standard Model and make testable predictions.

In this way, investigators are trying to reproduce the earlier success of finding why the chemical elements line up in the periodic table, representing the structure of atoms. Once chemists had gleaned this structure, they predicted new elements and elemental properties. Likewise, particle physicists today are trying to find out why the weight diagram of the Standard Model has the pattern it does, and once they do, they will be able to make predictions for what properties the particles should have and what new particles might exist.

The first attempt at such a theory was proposed in 1973, by Howard Georgi and Sheldon Glashow [see "A Unified Theory of Elementary Particles and Forces," by Howard Georgi; SCIENTIFIC AMERICAN, April 1981]. They found that the combined Lie group of the Standard Model fits snugly into the Lie group SU(5) as a subgroup. This SU(5) GUT made some distinctive predictions. First, fermions should have exactly the hypercharges that they do—a highly nontrivial success. Second, the weak mixing angle should be 38 degrees, in fair agreement with experiments. And finally, in addition to the 12 Standard Model bosons, there are 12 new force particles in SU(5), called *X* bosons.

It was the *X* bosons that got the theory into trouble. These new particles would allow protons to decay into lighter particles, which they cannot do in the Standard Model. In impressive experiments, including the observation of 50,000 tons of water in a converted Japanese mine, the predicted proton

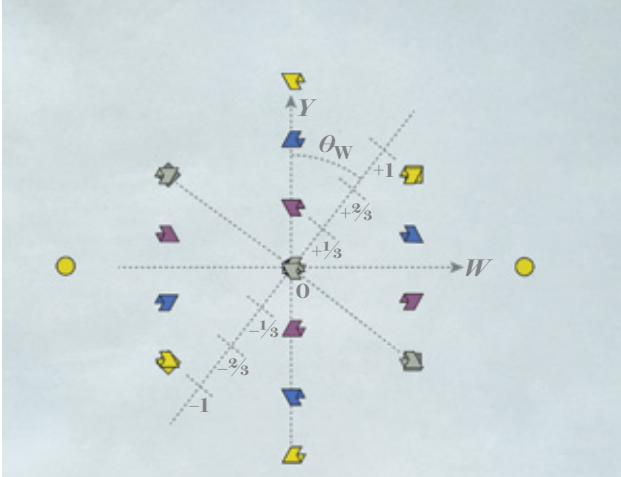
DIAGRAMS OF PARTICLE CHARGES

A Progression of Theories

Each type of elementary particle corresponds to a different fiber. These fibers twist around the various circular fibers of photons and other force particles; the twisting corresponds to their electric and other charges, which mathematicians call "weights." Each force of nature has its own type of charge and is mediated by its own force particles. The patterns of particles' charges describe the geometry of how the fibers twist around one another, governing how the corresponding particles can interact.

BOSONS	FERMIONS	
Photon	Neutrino	Left-handed spin-up
Weak bosons	Electron	Left-handed spin-down
Gluons	Up quark	Right-handed spin-up
Gravitons	Down quark	Right-handed spin-down
Frame-Higgs	Up quark	Right-handed spin-up anti
Weaker bosons	Down quark	Right-handed spin-down anti
X bosons	Up quark	Left-handed spin-up anti
More Higgs	Down quark	Left-handed spin-down anti

Particles come in two broad types, bosons (which transmit forces) and fermions (which constitute matter). Each fermion can come in several varieties: particle or antiparticle, left- or right-handed, spin up or down, and, for quarks, one of three colors. Every particle, identified by its charges, can be plotted in a weight diagram.



Electroweak Theory. Electromagnetism and the weak nuclear force emerge from the unified electroweak force. On its weight diagram, physicists plot particles by their hypercharge (*Y*) and weak nuclear charge (*W*). Higgs bosons (*partially hidden gray squares*) lie on an oblique line running from the top left to the bottom right, defining zero electric charge. Electrically charged particles lie on parallel lines. In this way, electric charge is a specific amalgam of hypercharge and weak charge.

© 2014 Scientific American

decay was not seen. Thus, physicists have ruled out this theory.

Despite the SU(5) theory's failures, its successes suggest that theorists are generally on the right track. A related Grand Unified Theory, developed around the same time, is based on the Lie group Spin(10). It produces the same hypercharges and weak mixing angle as SU(5) and predicts the existence of a new force, very similar to the weak force. This new "weaker" force, mediated by relatives of the weak-force bosons called W'^{+}, W'^{-} and W'^{3}, interacts with right-handed fermions, restoring left-right symmetry to the universe at short distances. Although this theory predicts an abundance of X bosons—a full 30 of them—it also indicates that proton decay would occur at a lower rate than for the SU(5) theory. So the theory remains viable.

Drawn a certain way, the weight diagram for the Spin(10) GUT shows that particle charges align in four concentric circles—an unusually pretty pattern [*see left panel on next page*]. The balance evident in this diagram arises for a deep reason: the Spin(10) Lie group with its 45 bosons, along with its representations of 16 fermions and 16 antifermions, comprises in fact all parts of a single Lie group, a special one known as the exceptional Lie group E6.

The exceptional groups play an exalted role in mathematics. Because there are only so many ways circles can twist around one another, there are only a handful of different kinds of Lie groups. Mathematicians completed their classification a century ago. We have already met two, SU and Spin, encountered quite often in physics. And among the Lie groups there are five exceptional cases that stand out: G2, F4, E6, E7 and E8. These Lie groups have especially intricate structures and deep connections to many areas of mathematics.

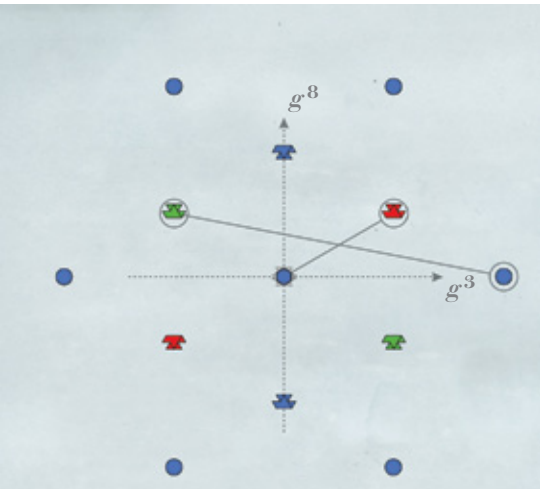
The fact that the bosons and fermions of Spin(10) and the Standard Model tightly fit the structure of E6, with its 78 generators, is remarkable. It provokes a radical thought. Up until now, physicists have thought of bosons and fermions as completely different. Bosons are parts of Lie group force fibers, and fermions are different kinds of fibers, twisting around the Lie groups. But what if bosons and fermions are parts of a single fiber? That is what the embedding of the Spin(10) GUT in E6 suggests. The structure of E6 includes both types of particles. In a radical unification of forces and matter, bosons and fermions can be combined as parts of a superconnection field.

Although several people have criticized this idea because it combines fermions and bosons in a way that at first appears fundamentally inconsistent, it relies on solid mathematics. And the curvature of this superconnection, describing the twisting of E6 over spacetime, succinctly describes the dynamics and interactions of bosons and fermions in the Standard Model. But E6 does not include the Higgs bosons or gravity.

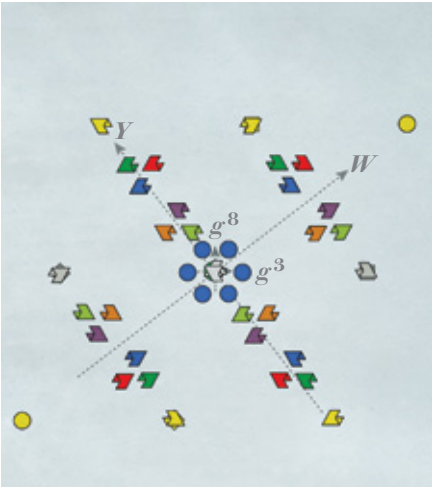
TAKING GRAVITY FOR A SPIN

ALBERT EINSTEIN originally described gravity as the curvature of spacetime. His mathematical machinery was state-of-the-art at the time, but researchers have gradually adopted a more modern, equivalent description of gravity based on a fiber bundle.

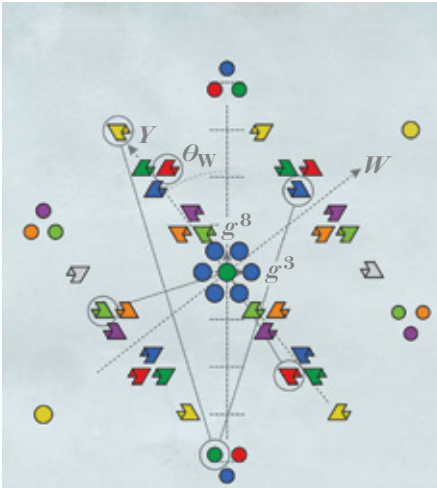
At every spacetime point, we can imagine three perpendicular rulers and a clock, called a frame of reference. Without the frame, spacetime would not be "spacetime" but just a four-dimensional fabric with no sense of orientation or distance. As we move to different points in spacetime, there are different sets of rulers and clocks, related to our original frame by a rotation. This rotation can be an ordinary rotation in space or, because Einstein showed that space and time are unified, a rotation of space into time. How the frame rotates from point to point is determined by the



Strong Nuclear Force. The eight gluons (six in a hexagon, two overlapping in the middle) that mediate the strong force interact with one another, quarks and antiquarks according to their strong-force charges (g^3 and g^8). Three quarks, one of each color, sit in a triangular pattern. When particles interact, such as the circled gluon interacting with a green quark to make a red quark, charges must balance.



Standard Model. Combining the weight diagrams of the electroweak and strong forces yields the Standard Model of particle physics—the prevailing theory of the natural world. The full four-dimensional diagram is flattened to fit on the page. All nongravitational particle interactions correspond to balancing charges in this diagram. A major goal of contemporary physics is to explain this pattern.



Grand Unified Theory. The electroweak and strong forces might be parts of a larger structure, such as the fiber shape called SU(5), with a pattern shown here. This shape has 12 new particles called X bosons, which allow a proton—two up quarks and a down quark (*circled*)—to change into a positron (antielectron) and a pion (up quark and anti-up-quark pair).

© 2014 Scientific American

spin connection, more commonly known as the gravitational field. The Lie group of possible rotations in three space directions and one time direction is Spin(1,3)—the Lie group of gravity. We feel the force of gravity because the gravitational spin connection field is rotating our frame as we move through time, attempting to steer us toward Earth's center.

Just as particles have different kinds of charge describing how they interact with Standard Model forces, they have a type of charge describing how they behave within space. Consider what happens if we rotate a ruler in space by 360 degrees: it returns to its original state. This ruler—and the gravitational frame field—has spatial spin charge of +1 or –1. But if we rotate a fermion, such as an electron, in space by 360 degrees, it does not return to the same state it started in. To return it to its original state, we have to rotate it by 720 degrees. The fermion has spin charge of ±½.

Spin charge plays a role in gravity because gravity, through the frame and spin connection, is related to the geometry of spacetime. As we did for the other forces, we can make a weight diagram for gravity based on spin [*see center panel below*]. A particle's spatial spin charge is its internal angular momentum, and its temporal spin charge is related to its motion through space. Fermions whose spatial spin and motion align, plotted in the upper right or lower left of the diagram, make a right-handed corkscrew as they travel through space. Fermions with opposite motion and spatial spin are left-handed.

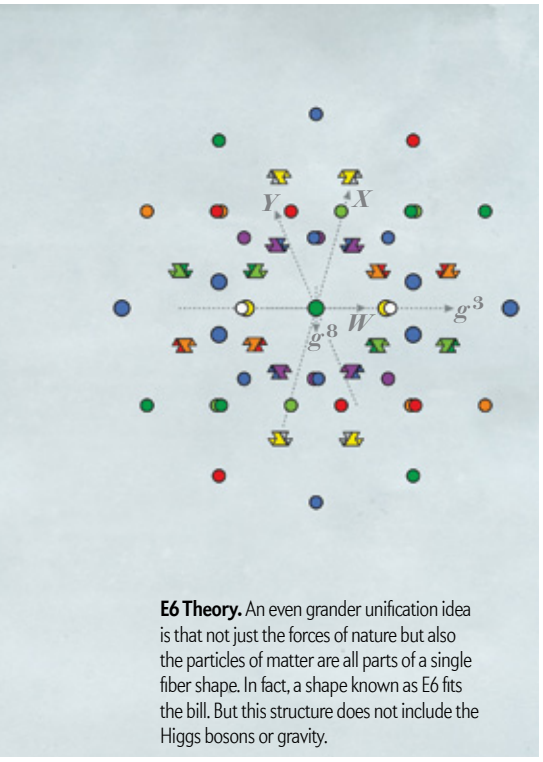
What is strange is that spin charge also has an unexpected relevance to the weak nuclear force. Only left-handed particles and right-handed antiparticles have weak charge and interact with the weak force. The fact that the weak force is sensitive to spin charge suggests that gravity and the other forces, though outwardly dissimilar, in fact are deeply related.

E PLURIBUS UNUM

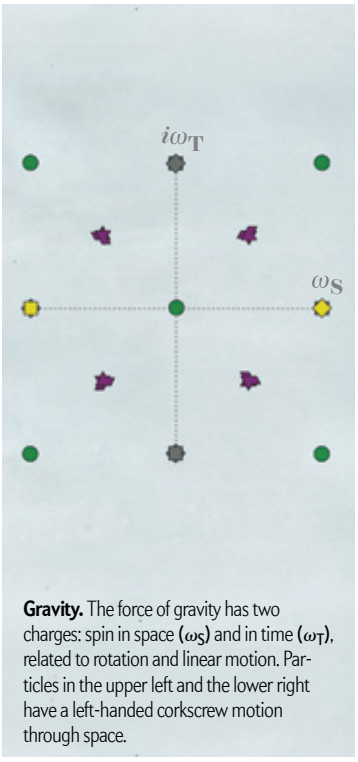
NOW IT IS JUST A MATTER of putting the pieces together. With gravity described by Spin(1,3) and the favored Grand Unified Theory based on Spin(10), it is natural to combine them using a single Lie group, Spin(11,3), yielding a Gravitational Grand Unified Theory—as introduced in 2009 by Roberto Percacci of the International School for Advanced Studies in Trieste and Fabrizio Nesti, then at the University of Ferrara in Italy. It brings us close to a full Theory of Everything.

The Spin(11,3) Lie group allows for blocks of 64 fermions and, amazingly, predicts their spin, electroweak and strong charges perfectly. It also automatically includes a set of Higgs bosons and the gravitational frame; in fact, they are unified as "frame-Higgs" generators in Spin(11,3). The curvature of the Spin(11,3) fiber bundle correctly describes the dynamics of gravity, the other forces and the Higgs. It even includes a cosmological constant that explains cosmic dark energy. Everything falls into place.

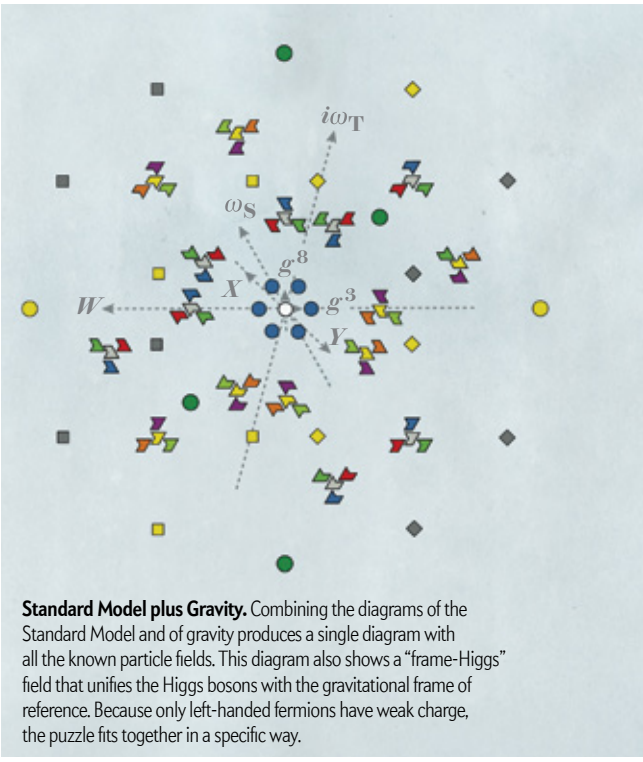
Skeptics objected that such a theory should be impossible. It appears to violate a theorem in particle physics, the Coleman-Mandula theorem, which forbids combining gravity with the other forces in a single Lie group. But the theorem has an important loophole: it applies only when spacetime exists. In the Spin(11,3) theory (and in E8 theory), gravity is unified with the other forces only before the full Lie group symmetry is broken, and when that is true, spacetime does not yet exist. Our universe begins when the symmetry breaks: the frame-Higgs field becomes nonzero, singling out a specific direction in the unifying Lie group. At this instant, gravity becomes an independent force, and spacetime comes into existence with a bang. Thus, the theorem is always satisfied. The dawn of time was the breaking of perfect symmetry.



E6 Theory. An even grander unification idea is that not just the forces of nature but also the particles of matter are all parts of a single fiber shape. In fact, a shape known as E6 fits the bill. But this structure does not include the Higgs bosons or gravity.



Gravity. The force of gravity has two charges: spin in space (ω_S) and in time (ω_T), related to rotation and linear motion. Particles in the upper left and the lower right have a left-handed corkscrew motion through space.



Standard Model plus Gravity. Combining the diagrams of the Standard Model and of gravity produces a single diagram with all the known particle fields. This diagram also shows a "frame-Higgs" field that unifies the Higgs bosons with the gravitational frame of reference. Because only left-handed fermions have weak charge, the puzzle fits together in a specific way.

© 2014 Scientific American

The weight diagram of the Spin(11,3) theory is finely patterned and balanced. Its symmetry, like that of the Spin(10) GUT, hints at deeper, exceptional mathematics. This elegant pattern of particles is part of what is perhaps the most beautiful structure in all of mathematics, the largest simple exceptional Lie group, E8. Just as E6 contains the structure of the Spin(10) Grand Unified Theory, with its 16 fermions, the E8 Lie group contains the structure of the Spin(11,3) Gravitational Grand Unified Theory, with its 64 Standard Model fermions, including their spins. In this way, gravity and the other known forces, the Higgs, and one generation of Standard Model fermions are all parts of the unified superconnection field of an E8 fiber bundle.

The E8 Lie group, with 248 generators, has a wonderfully intricate structure. In addition to gravity and the Standard Model particles, E8 includes W', Z' and X bosons, a rich set of Higgs bosons, novel particles called mirror fermions, and axions—a cosmic dark matter candidate. Even more intriguing is a symmetry of E8 called triality. Using triality, the 64 generators of one generation of Standard Model fermions can be related to two other blocks of 64 generators. These three blocks might intermix to reproduce the three generations of known fermions.

COLLIDING WITH REALITY

IN THIS WAY, the physical universe could emerge naturally from a mathematical structure without peer. The theory tells us what Higgs bosons are, how gravity and the other forces emerge from symmetry-breaking, why fermions exist with the spins and charges they have, and why all these particles interact as they do. Although this new theory continues to be promising, much work remains to be done. We need to figure out how three generations of fermions unfold, how they mix and interact with the Higgs to get their masses, and exactly how E8 theory works within the context of quantum theory.

If E8 theory is correct, it is likely the Large Hadron Collider will detect some of its predicted particles. If, on the other hand, the collider detects new particles that do not fit E8's pattern, that could be a fatal blow for the theory. In either case, any new particles will take their place in a weight diagram, leading us toward some geometric structure at the heart of nature. And if the structure of the universe at the scales of elementary particles does turn out to be described by E8, with its 248 sets of circles wrapping around one another, twisting and dancing over spacetime in all possible ways, then we will have achieved a complete unification and have the satisfaction of knowing we live in an exceptionally beautiful universe. SA

A. Garrett Lisi balances his time between research in theoretical physics and surfing. As an itinerant scientist, he is in the process of realizing a lifelong dream: founding the Pacific Science Institute, located on the Hawaiian island of Maui.

James Owen Weatherall is an assistant professor of logic and the philosophy of science at the University of California, Irvine. His 2013 book *The Physics of Wall Street* explores the history of how ideas have moved from physics into financial modeling.

MORE TO EXPLORE

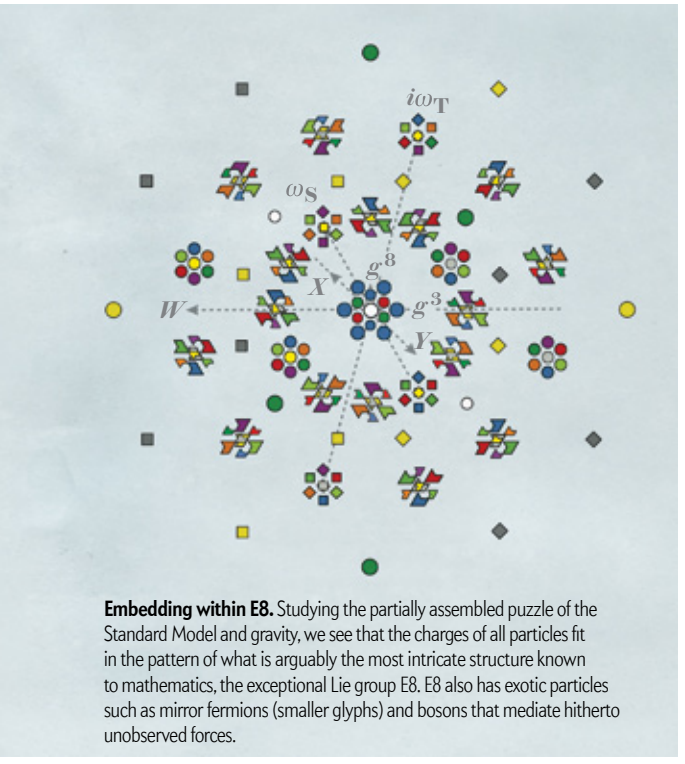
Deep Down Things. Bruce A. Schumm. Johns Hopkins University Press, 2004.

Lie Groups, Physics, and Geometry. Robert Gilmore. Cambridge University Press, 2008.

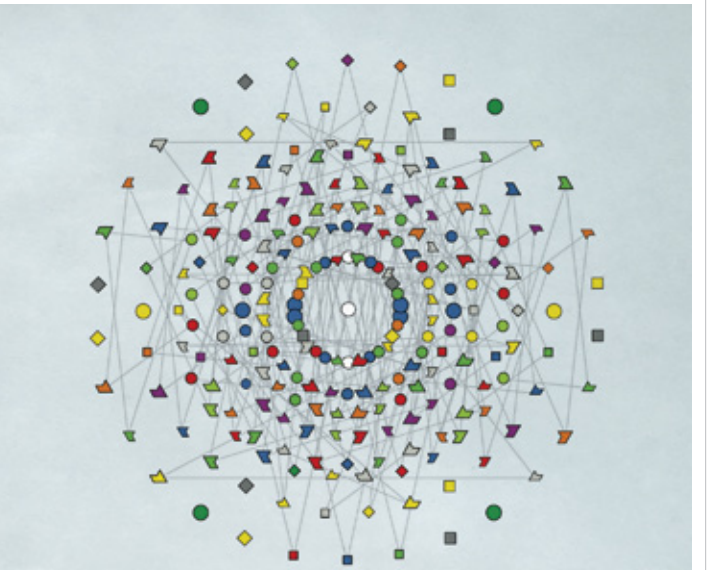
The Beauty of Particle Physics. A. Garrett Lisi. www.ted.com/index.php/talks/garrett_lisi_on_his_theory_of_everything.html

An Exceptionally Simple Theory of Everything. A. Garrett Lisi. http://arxiv.org/abs/0711.0770

scientificamerican.com/magazine/sa

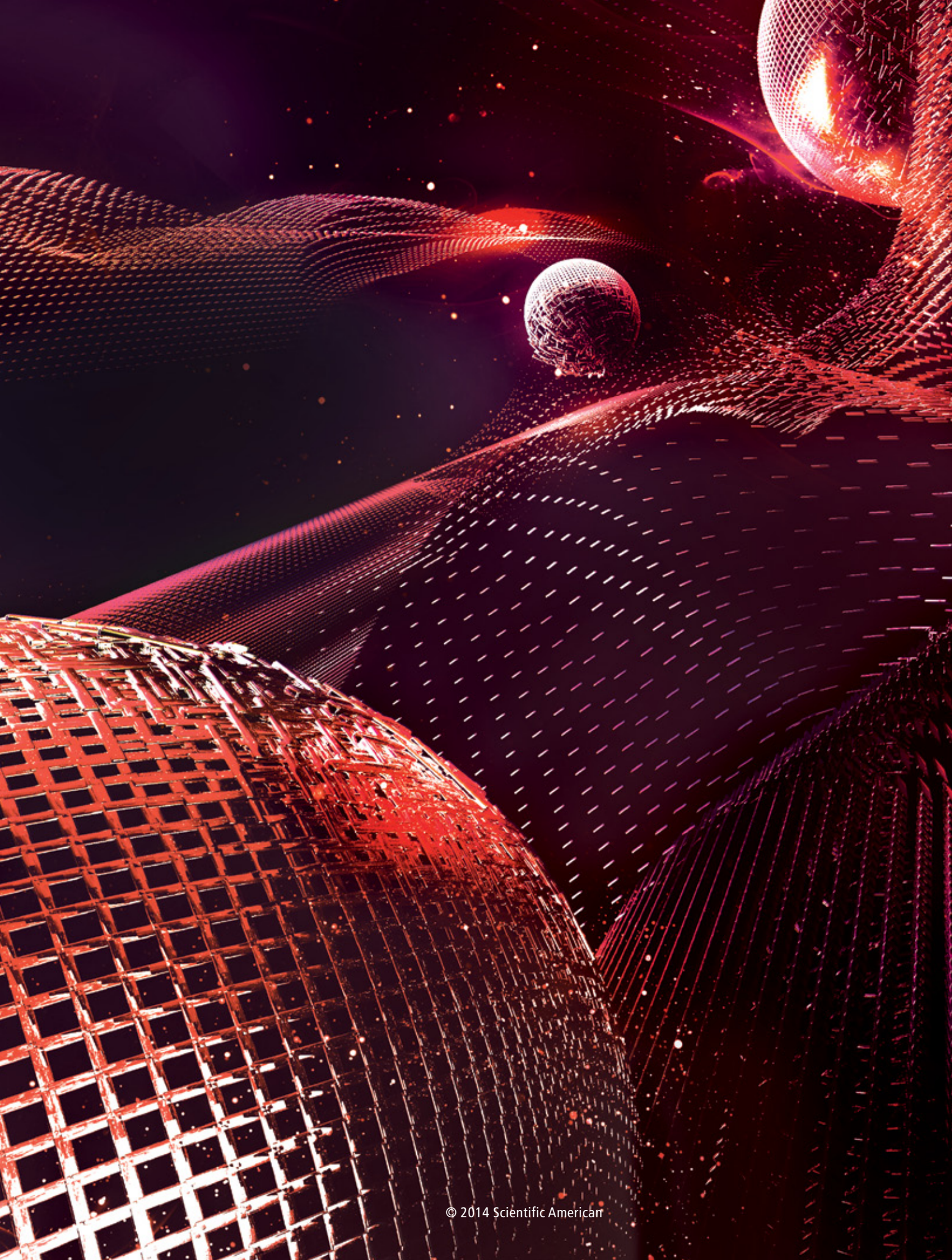


Embedding within E8. Studying the partially assembled puzzle of the Standard Model and gravity, we see that the charges of all particles fit in the pattern of what is arguably the most intricate structure known to mathematics, the exceptional Lie group E8. E8 also has exotic particles such as mirror fermions (smaller glyphs) and bosons that mediate hitherto unobserved forces.



E8 Theory. The embedding within E8 suggests that every fiber there is—every force, every known particle of matter and a clutch of additional particles that might account for cosmic dark matter—could be parts of this one exquisite shape. E8 even has a special symmetry called triality that relates its parts, which might explain why fermions come in three progressively heavier varieties called generations. E8 theory may be the long-sought Theory of Everything.

© 2014 Scientific American



© 2014 Scientific American

EXPERIMENTAL PHYSICS

IS SPACE DIGITAL?

An experiment going up outside of Chicago will attempt to measure the intimate connections among information, matter and spacetime. If it works, it could rewrite the rules for 21st-century physics

By Michael Moyer

Illustration by Vault49

© 2014 Scientific American

CRAIG HOGAN BELIEVES THAT THE WORLD IS FUZZY. THIS IS NOT A METAPHOR. HOGAN, A physicist at the University of Chicago and director of the Fermilab Particle Astrophysics Center near Batavia, Ill., thinks that if we were to peer down at the tiniest subdivisions of space and time, we would find a universe filled with an intrinsic jitter, the busy hum of static. This hum comes not from particles bouncing in and out of being or other kinds of quantum froth that physicists have argued about in the past. Rather Hogan's noise would come about if space was not, as we have long assumed, smooth and continuous, a glassy backdrop to the dance of fields and particles. Hogan's noise arises if space is made of chunks. Blocks. Bits. Hogan's noise would imply that the universe is digital.

It is a breezy, early autumn afternoon when Hogan takes me to see the machine he is building to pick out this noise. A bright-blue shed rises out of the khaki prairie of the Fermilab campus, the only sign of new construction at this nearly half-century-old facility. A fist-wide pipe runs 40 meters from the shed to a long, perpendicular bunker, the former home of a beam that for decades shot subatomic particles north toward Minnesota. The bunker has been reclaimed by what Hogan calls his Holometer, a device designed to amplify the jitter in the fabric of space.

He pulls out a thick piece of sidewalk chalk and begins to write on the side of the cerulean shed, his impromptu lecture detailing how a few lasers bouncing through the tubes can amplify the fine-grain structure of space. He begins by explaining how the two most successful theories of the 20th century—quantum mechanics and general relativity—cannot possibly be reconciled. At the smallest scales, both break down into gibberish. Yet this same scale seems to be special for another reason: it happens to be intimately connected to the science of information—the 0's and 1's of the universe. Physicists have, over the past couple of decades, uncovered profound insights into how the universe stores information—even going so far as to suggest that information, not matter and energy, constitutes the most basic unit of existence. Information rides on tiny bits; from these bits comes the cosmos.

If we take this line of thinking seriously, Hogan says, we should be able to measure the digital noise of space. Thus, he has devised an experiment to explore the buzzing at the universe's most fundamental scales. He will be the first to tell you that it might not work—that he may see nothing at all. His effort is an experiment in the truest sense—a trial, a probe into the unknown. "You cannot take the well-tested physics of spacetime and the well-tested physics of quantum mechanics and calculate what we'll see," Hogan says. "But to me, that's the reason to do the experiment—to go in and see."

And if he does see this jitter? Space and time are not what we thought. "It changes the architecture of physics," Hogan says.

FOR MANY YEARS PARTICLE PHYSICS HAS NOT OPERATED on this sort of exploratory model. Scientists spent the late 1960s and early 1970s developing a web of theories and insights that we now know as the Standard Model of particle physics. In the decades since, experiments have tested it with increasing depth and precision. "The pattern has been that the theory community has come up with an idea—for example, the Higgs boson—and you have a model. And the model makes a prediction, and the experiment rules it out or not," Hogan says. Theory comes first, experiments later.

This conservatism exists for a very good reason: particle physics experiments can be outrageously expensive. The Large Hadron Collider (LHC) at CERN near Geneva required around $5 billion to assemble and currently occupies the attention of thousands of physicists around the world. It is the most sophisticated, complex and precise machine ever built. Scientists openly wonder if the next generation of particle collider—at higher energies, larger sizes and greater expenses—will prove too ambitious. Humanity may simply refuse to pay for it.

IN BRIEF

Space may not be smooth and continuous. Instead it may be digital, composed of tiny bits. Physicists have assumed that these bits are far too small to measure with current technology.

Yet one scientist thinks that he has devised a way to detect the bitlike structure of space. His machine—at present under construction—will attempt to measure its grainy nature.

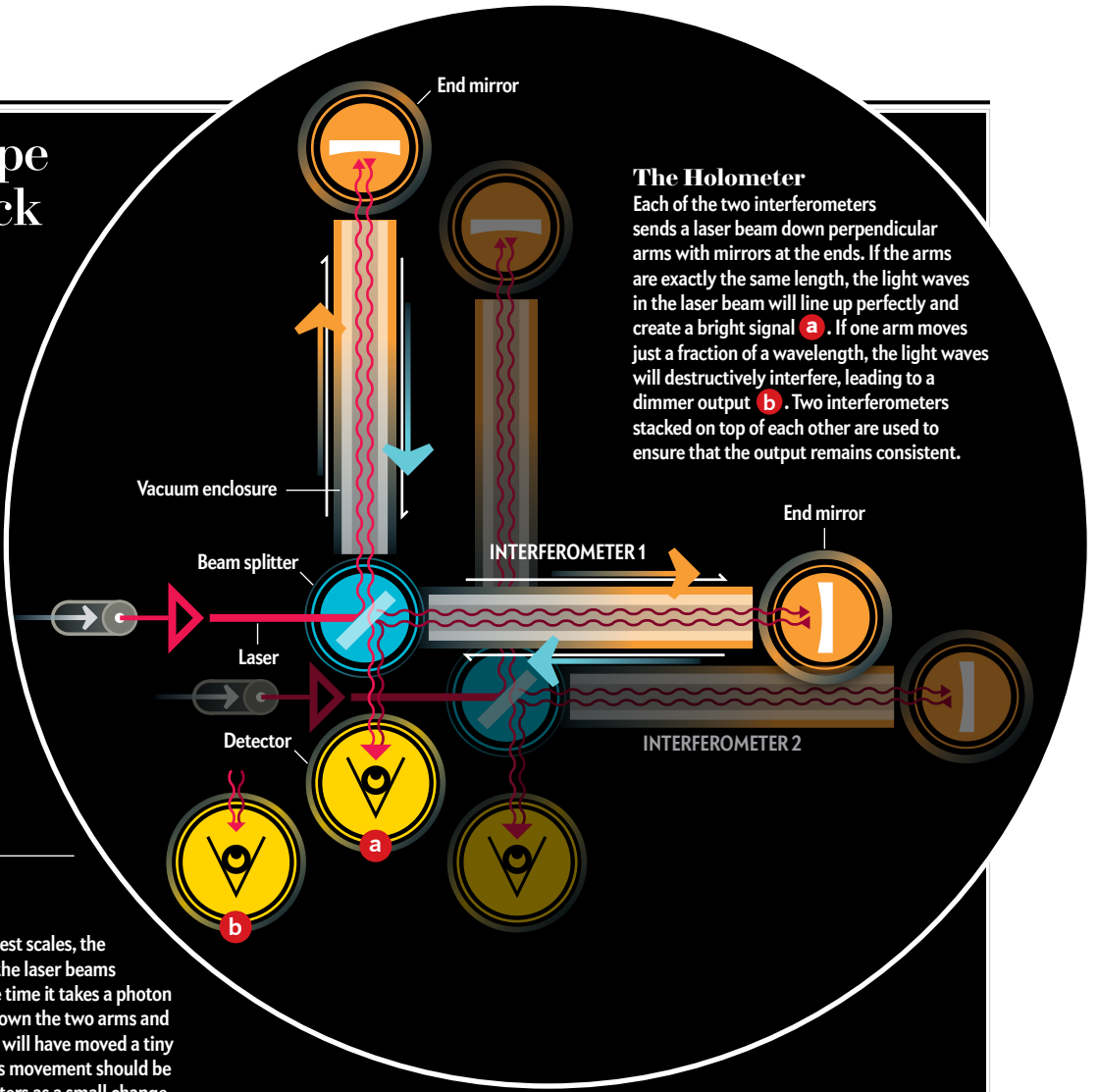
The experiment is one of the first to investigate the principle that the universe emerges from information—specifically, information that is imprinted on two-dimensional sheets.

If successful, the experiment will shift the foundations of what we know about space and time, providing a glimpse of a new physics that could supplant our existing understanding.

© 2014 Scientific American

A Microscope to the Planck Length

With his Holometer, Craig Hogan will try to measure a fundamental jitter in spacetime at the smallest scale. The device consists of two interferometers, instruments that amplify very small changes in distance (*right*). Detecting a jitter would indicate that spacetime is digital—divided into discrete packets (*bottom*).

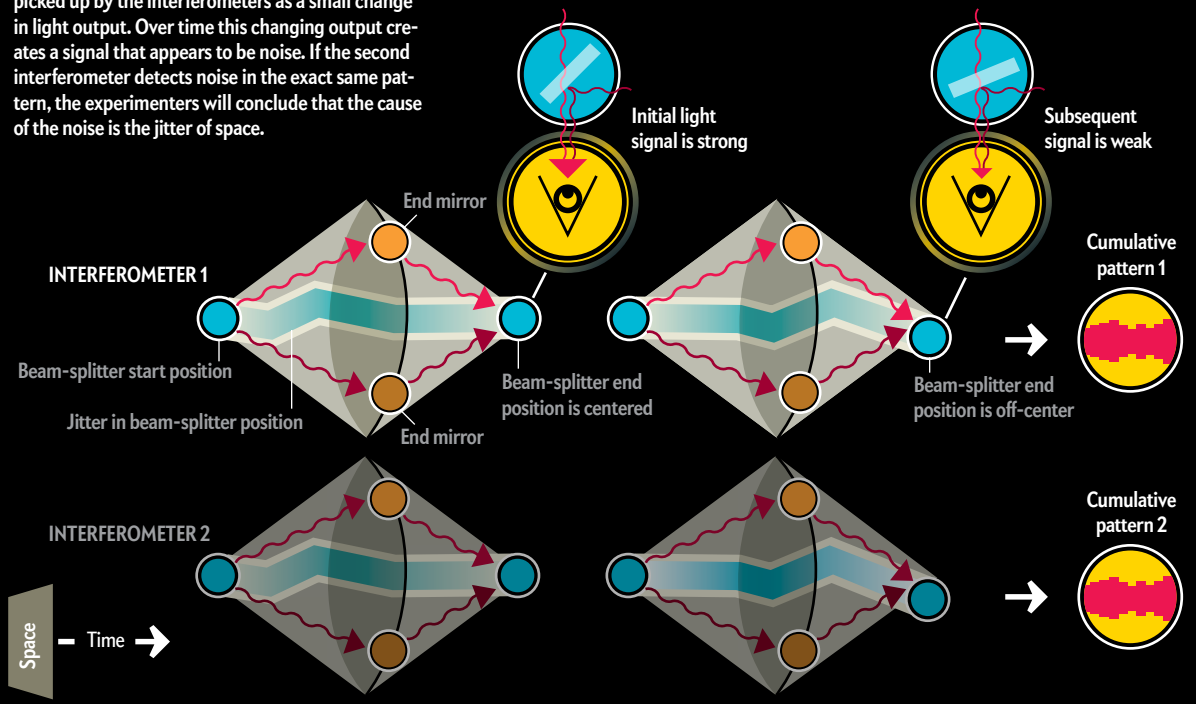


The Holometer

Each of the two interferometers sends a laser beam down perpendicular arms with mirrors at the ends. If the arms are exactly the same length, the light waves in the laser beam will line up perfectly and create a bright signal **a**. If one arm moves just a fraction of a wavelength, the light waves will destructively interfere, leading to a dimmer output **b**. Two interferometers stacked on top of each other are used to ensure that the output remains consistent.

The Jitter of Space

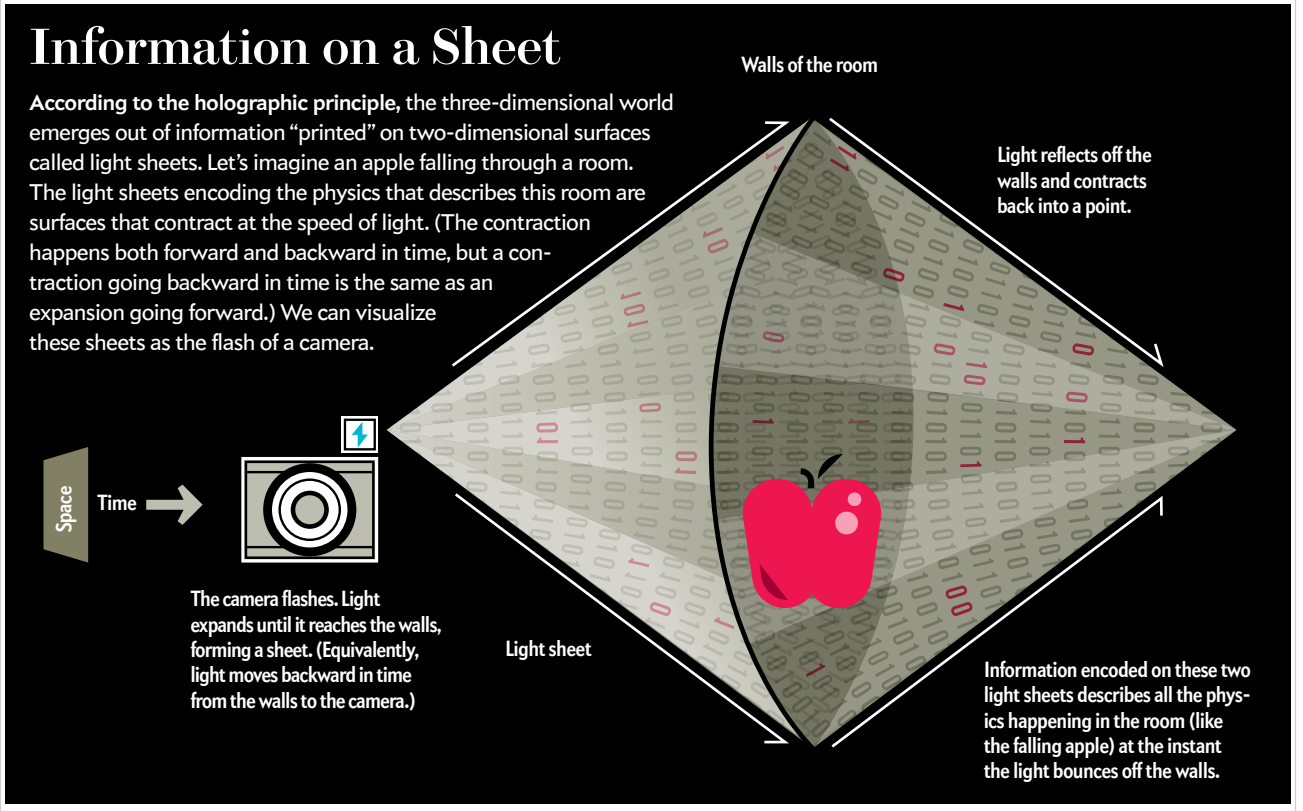
If space is frothy on the smallest scales, the beam splitter that separates the laser beams should bounce around. In the time it takes a photon to travel out from the laser, down the two arms and back again, the beam splitter will have moved a tiny bit in a random direction. This movement should be picked up by the interferometers as a small change in light output. Over time this changing output creates a signal that appears to be noise. If the second interferometer detects noise in the exact same pattern, the experimenters will conclude that the cause of the noise is the jitter of space.



Illustrations by Leandro Castelao

© 2014 Scientific American

THE HOLOGRAPHIC PRINCIPLE


Information on a Sheet
According to the holographic principle, the three-dimensional world emerges out of information "printed" on two-dimensional surfaces called light sheets. Let's imagine an apple falling through a room. The light sheets encoding the physics that describes this room are surfaces that contract at the speed of light. (The contraction happens both forward and backward in time, but a contraction going backward in time is the same as an expansion going forward.) We can visualize these sheets as the flash of a camera.
Walls of the room
Light reflects off the walls and contracts back into a point.
Space
Time
The camera flashes. Light expands until it reaches the walls, forming a sheet. (Equivalently, light moves backward in time from the walls to the camera.)
Light sheet
Information encoded on these two light sheets describes all the physics happening in the room (like the falling apple) at the instant the light bounces off the walls.

A typical experiment at the LHC might include more than 3,000 researchers. At Fermilab, Hogan has assembled a loosely knit team of 20 or so, a figure that includes senior advisers at the Massachusetts Institute of Technology and the University of Michigan who do not participate in day-to-day work at the site. Hogan is primarily a theoretical physicist—largely unfamiliar with the vagaries of vacuum pumps and solid-state lasers—and so he has enlisted as co-leader Aaron Chou, an experimentalist who happened to arrive at Fermilab at about the same time Hogan was putting his proposal forward. In 2011 they were awarded $2 million, which at the LHC would buy you a superconducting magnet and a cup of coffee. The money will fund the entire project. "We don't do any high-tech thing if low-tech will do," Hogan says.

The experiment is so cheap because it is basically an update of the experiment that so famously destroyed the 19th century's established wisdom about the backdrop of existence. By the early 1800s physicists knew that light behaved as a wave. And waves, scientists knew. From a ripple in a pond to sound moving through the air, all waves seemed to share a few essential features. Like sculptures, waves always require a medium—some physical substrate that the waves must travel through. Because light is a wave, the thinking went, it must also require a medium, an invisible substance that permeated the universe. Scientists called this hidden medium the ether.

In 1887 Albert Michelson and Edward Morley designed an experiment that would search for this ether. They set up an interferometer—a device with two arms in the shape of an L that was optimized to measure change. A single source of light would travel the length of both arms, bounce off mirrors at the ends, then recombine where it began. If the length of time it took the light to travel down either arm changed by even a fraction of a microsecond, the recombined light would glow darker. Michelson and Morley set up their interferometer and monitored the light for months as the earth moved around the sun. Depending on which way the earth was traveling, the stationary ether should have altered the time it took for the light to bounce down the perpendicular arms. Measure this change, and you have found the ether.

Of course, the experiment found no such thing, thus beginning the destruction of a cosmology hundreds of years old. Yet like a forest obliterated by fire, clearing the ether made it possible for revolutionary new ideas to flourish. Without an ether, light traveled the same speed no matter how you were moving. Decades later Albert Einstein seized this insight to derive his theories of relativity.

Hogan's interferometer will search for a backdrop that is much like the ether—an invisible (and possibly imaginary) substrate that permeates the universe. By using two Michelson interferometers stacked on top of each other, he intends to probe the smallest scales in the universe, the distance at which both quantum mechanics and relativity break down—the region where information lives as bits.

THE PLANCK SCALE IS NOT JUST SMALL—IT IS THE smallest. if you took a particle and confined it inside a cube less than one Planck length on each side, general relativity says that it would weigh more than a black hole of that same size. But the laws of quantum mechanics say that any black hole smaller

© 2014 Scientific American

than a Planck length must have less than a single quantum of energy, which is impossible. At the Planck length lies paradox.

Yet the Planck length is much more than the space where quantum mechanics and relativity fall apart. In the past few decades an argument over the nature of black holes has revealed a wholly new understanding of the Planck scale. Our best theories may break down there, but in their place something else emerges. The essence of the universe is information, so this line of thinking goes, and the fundamental bits of information that give rise to the universe live on the Planck scale.

As explained by Stanford University physicist Leonard Susskind during a public lecture in 2011, "information" is about distinctions. "It is a very basic principle of physics that distinctions never disappear," he said. "They might get scrambled or all mixed up, but they never go away." Even after this magazine gets dissolved into pulp at the recycling plant, the information on these pages will be reorganized, not eliminated. In theory, the decay can be reversed—the pulp reconstructed into words and photographs—even if, in practice, the task appears impossible.

SPACE DOES NOT SIT STILL, A SMOOTH BACKDROP TO THE COSMOS. INSTEAD QUANTUM FLUCTUATIONS MAKE SPACE BRISTLE AND VIBRATE, SHIFTING THE WORLD ALONG WITH IT.

Physicists have long agreed on this principle except in one special case. What if this magazine were to be thrown into a black hole? Nothing can ever emerge from a black hole, after all. Throw these pages into a black hole, and that black hole will appear almost exactly the same as it did before—just a few grams heavier, perhaps. Even after Stephen Hawking showed in 1975 that black holes can radiate away matter and energy (in the form that we now call Hawking radiation), this radiation seemed to be devoid of structure, a flat bleat at the cosmos. He thus concluded that black holes must destroy information.

Nonsense, argued a number of Hawking's colleagues, among them Susskind and Gerard 't Hooft, a theoretical physicist at Utrecht University in the Netherlands who would go on to win the Nobel Prize in 1999. "The whole structure of everything we know would disintegrate if you opened the door even a tiny bit for the notion of information to be lost," Susskind says.

Hawking was not easily convinced, however, and so over the following two decades physicists developed a new theory that could account for the discrepancy. This is the holographic principle, and it holds that when an object falls into a black hole, the stuff inside may be lost, but the object's information is somehow imprinted onto a surface around the black hole. With the right tools, you could theoretically reconstruct this magazine from a black hole just as you could from the pulp at the recycling plant. The black hole's event horizon—the point of no return—serves double duty as a ledger. Information is not lost.

The principle is more than just an accounting trick. It implies that whereas the world we see around us appears to take place in three dimensions, all the information about it is stored on surfaces that have just two dimensions [see "Information in the Holographic Universe," by Jacob D. Bekenstein; Scientific American, August 2003]. What is more, there is a limit to how much information can be stored on a given surface area. If you divide a surface up like a checkerboard, each square two Planck lengths on a side, the information content will always be less than the number of squares.

In a series of papers in 1999 and 2000 Raphael Bousso, now at the University of California, Berkeley, showed how to extend this holographic principle beyond the simple surfaces around black holes. He imagined an object surrounded by flashbulbs popping off in the dark. Light that traveled inward defined a surface—a bubble collapsing at the speed of light. It is on this two-dimensional surface—the so-called light sheet—that all the information about you (or a flu virus or a supernova) is stored [*see box on opposite page*].

This light sheet, according to the holographic principle, does a lot of work. It contains information about the position of every particle inside the sheet, every electron and quark and neutrino, and every force that acts on them. Yet it would be wrong to think about the light sheet as a piece of film, passively recording the real stuff that happens out in the world. Instead the light sheet comes first. It projects the information contained on its surface out into the world, creating all that we see. In some interpretations, the light sheet does not just generate all the forces and particles—it gives rise to the fabric of spacetime itself. "I believe that spacetime is what we call emergent," says Herman Verlinde, a physicist at Princeton University and a former student of 't Hooft. "It will come out of a bunch of 0's and 1's."

One problem: although physicists mostly agree that the holographic principle is true—that information on nearby surfaces contains all the information about the world—they know not how the information is encoded, or how nature processes the 1's and 0's, or how the result of that processing gives rise to the world. They suspect the universe works like a computer—that information conjures up what we perceive to be physical reality—but right now that computer is a big black box.

Ultimately the reason why physicists are so excited about the holographic principle, the reason they spent decades developing it—other than convincing Hawking that he was mistaken, of course—is because it articulates a deep connection among information, matter and gravity. In the end, the holographic principle could reveal how to reconcile the two tremendously successful yet mutually incompatible pillars of 20th-century physics: quantum mechanics and general relativity. "The holographic principle is a signpost to quantum gravity," Bousso says, an observation that points the way toward a theory that will supersede our current understanding of the world. "We might need more signposts."

INTO ALL THIS CONFUSION COMES HOGAN, WITH NO grand theory of everything, armed with his simple Holometer. But Hogan does not need a grand theory. He does not have to solve all these difficult problems. All he has to do is figure out one fundamental fact: Is the universe a bitlike world, or isn't it? If he can do that, he will indeed have produced a signpost—a giant arrow pointing in the direction of a digital universe, and physicists would know which way to go.

According to Hogan, in a bitlike world, space is itself quantum—it emerges from the discrete, quantized bits at the Planck

© 2014 Scientific American

scale. And if it is quantum, it must suffer from the inherent uncertainties of quantum mechanics. It does not sit still, a smooth backdrop to the cosmos. Instead quantum fluctuations make space bristle and vibrate, shifting the world around with it. "Instead of the universe being this classical, transparent, crystalline-type ether," says Nicholas B. Suntzeff, an astronomer at Texas A&M University, "at a very, very small scale, there are these little foamlike fluctuations. It changes the texture of the universe tremendously."

The trick is getting down to the level of this spacetime foam and measuring it. And here we run into the problem of the Planck length. Hogan's Holometer is an attempt to flank a full-scale assault on the Planck length—a unit so small that measuring it with a conventional experiment (such as a particle accelerator) would involve building a machine the approximate size of the Milky Way.

Back when Michelson and Morley were investigating the (nonexistent) ether, their interferometer measured a tiny change—the change in the speed of light as the earth moved around the sun—by comparing two light beams that had traveled a reasonably long way. In effect, that distance multiplied the signal. So it is with Hogan's Holometer. His strategy for getting down to the Plank length is to measure the accumulated errors that accrue when dealing with any jittery quantum system.

"If I look at my TV set or my computer monitor, everything looks nice and smooth," Chou says. "But if you look at it close-up, you can see the pixels." As it would be with spacetime. At the level we humans are comfortable with—the scale of people and buildings and microscopes—space appears to be this smooth, continuous thing. We never see a car move down the street by instantaneously leaping from one place to the next as if lit by God's own strobe light.

Yet in Hogan's holographic world, this is exactly what happens. Space is itself discrete—or, in the parlance of our times, "quantized" [see "Atoms of Space and Time," by Lee Smolin; Scientific American, January 2004]. It emerges out of some deeper system, some fundamentally quantum system that we do not yet understand. "It's a slight cheat because I don't have a theory," Hogan says. "But it's only a first step. I can say to these gravitational theorists, 'You guys figure out how it works.'"

HOGAN'S HOLOMETER IS SET UP MUCH LIKE MICHELSON and Morley's, if Michelson and Morley had access to microelectronics and two-watt lasers. A laser hits a beam splitter that separates the light into two. These beams travel down the two 40-meter-long arms of an L-shaped interferometer, bounce off mirrors at each end, then return to the beam splitter and recombine. Yet instead of measuring the motion of the earth through the ether, Hogan is measuring any change in the length of the paths as a result of the beam splitter being jostled around on the fabric of space. If at the Planck scale, spacetime thrashes around like a roiling sea, the beam splitter is the dinghy pitching through the froth. In the time it takes the laser beams to travel out and back through the Holometer, the beam splitter will have jiggled just enough Planck lengths for its motion to be detected [*see box on page 107*].

Of course, you might imagine a lot of reasons why a beam splitter might move a few Planck lengths here and there—the rumbling of a car engine outside the building, for instance, or a stiff Illinois wind shaking the foundations.

Such concerns have bedeviled the scientists behind another interferometry project, the twin Laser Interferometer Gravitational-Wave Observatory (LIGO) detectors outside of Livingston, La., and Hanford, Wash. These massive experiments were built to observe gravitational waves—the ripples in spacetime that follow cosmic cataclysms such as neutron star collisions. Unfortunately for the LIGO scientists, gravitational waves shake the ground at the same frequency as other not so interesting things—passing trucks and falling trees, for instance. As such, the detectors have to be completely isolated against noise and vibration. (A proposed wind farm near the Hanford facility caused much consternation among physicists because the mere vibration of the blades would have swamped the detectors with noise.)

The shaking that Hogan is looking for happens much faster—a vibration that jitters back and forth a million times a second. As such, it is not subject to the same noise concerns—only the possible interference from nearby AM radio stations broadcasting at the same frequency. "Nothing moves at that frequency," says Stephan Meyer, a University of Chicago physicist and LIGO veteran who is working on the Holometer. "If we discover that it's moving anyway, that's one of the things that we'll take as a sure sign" that the jitter is real.

And in the world of particle physics, sure signs can be hard to come by. "This is old-fashioned in a way," Hogan says. "It appeals to this old-fashioned style of physics, which is, 'We're going to go and find out what nature does, without prejudice.'" To illustrate, he likes to tell a parable about the origins of relativity and quantum mechanics. Einstein invented the theory of general relativity by sitting at his desk and working out the mathematics from first principles. There were few experimental quandaries that it solved—indeed, its first real experimental test would not come for years. Quantum mechanics, on the other hand, was imposed on the theorists by the puzzling results of experiments. ("No theorist in his right mind would have invented quantum mechanics unless forced to by data," Hogan says.) Yet it has become the most successful theory in the history of science.

In the same way, theorists have for many years been building beautiful theories such as string theory, although it remains unclear how or if it can ever be tested. Hogan sees the purpose of his Holometer as a way to create the puzzling data that future theorists will have to explain. "Things have been stuck for a long time," he says. "How do you unstick things? Sometimes they get unstuck with an experiment." SA

Michael Moyer is a senior editor at *Scientific American*.

MORE TO EXPLORE

The Holographic Principle. Raphael Bousso in *Reviews of Modern Physics*, Vol. 74, No. 3, pages 825–874; 2002. http://arxiv.org/abs/hep-th/0203101

Information in the Holographic Universe. Jacob D. Bekenstein in *Scientific American*, Vol. 289, No. 2, pages 58–65; August 2003.

Atoms of Space and Time. Lee Smolin in *Scientific American*, Vol. 290, No. 1, pages 66–75; January 2004.

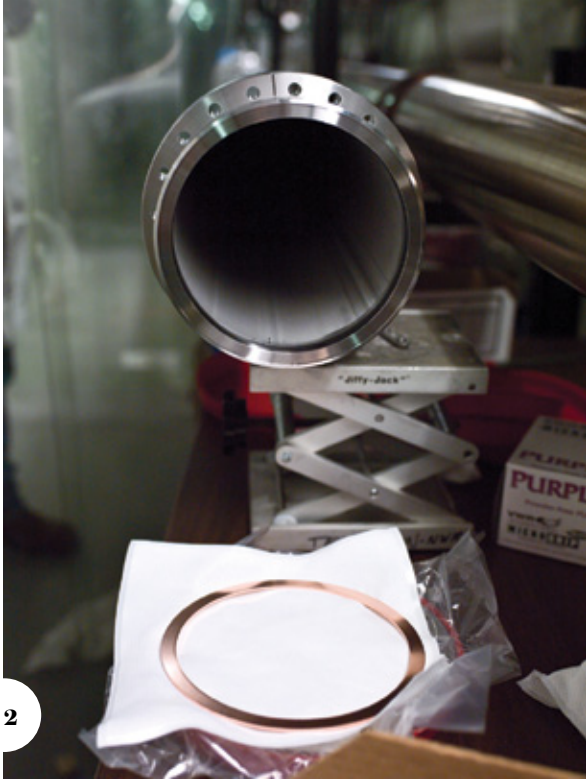
Now Playing: Reality. In 3-D. T. Kunz in *Symmetry*, Vol. 8, No. 3, pages 22–25; October 2011.

Interferometers as Probes of Planckian Quantum Geometry. Craig J. Hogan. http://arxiv.org/abs/1002.4880

scientificamerican.com/magazine/sa

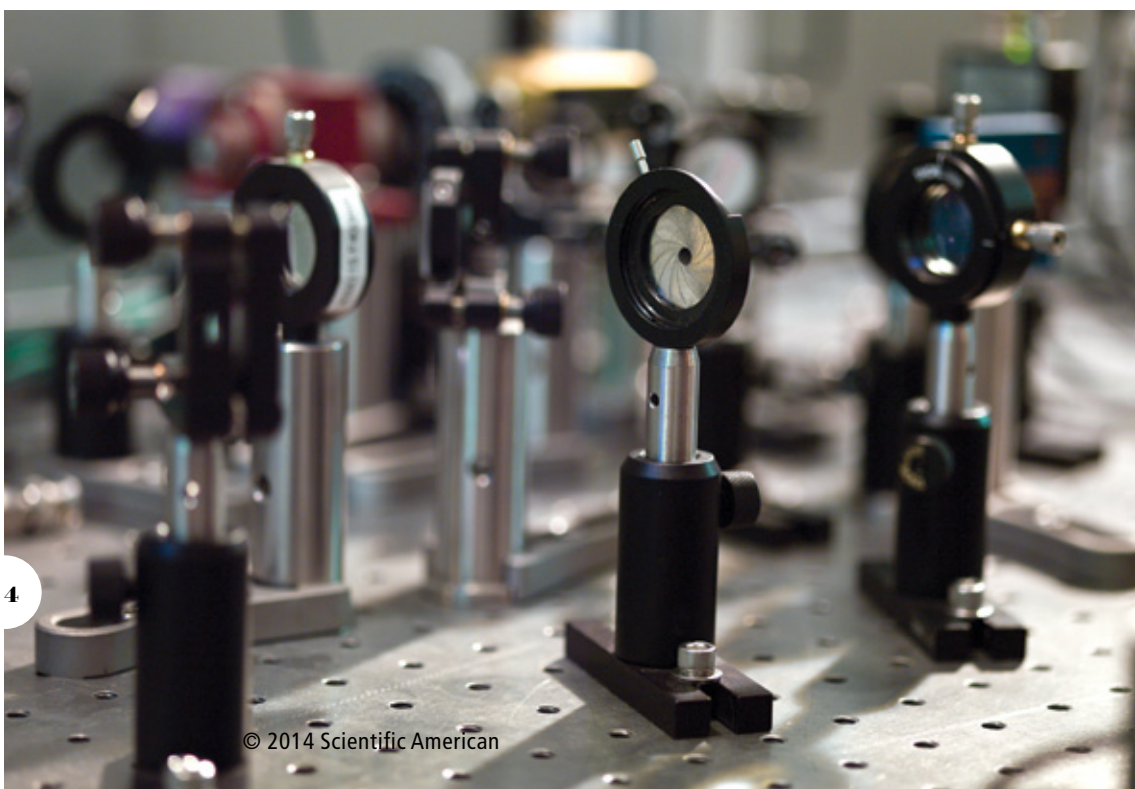
© 2014 Scientific American

1



2

3

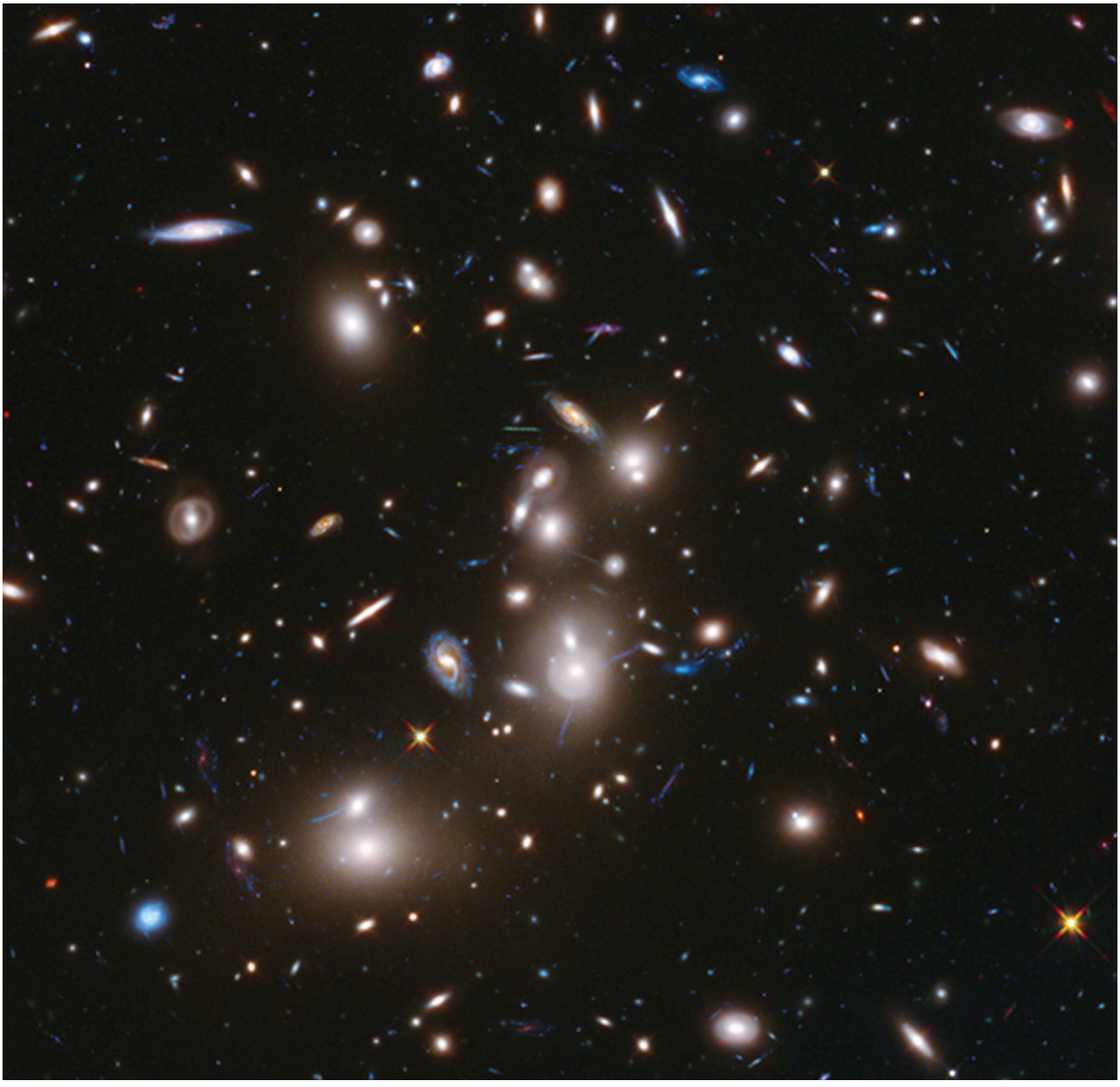


4

CRAIG HOGAN (*1*), director of the Fermilab Center for Particle Astrophysics, pauses in his office. Hogan and his team are building the Holometer at a site about a kilometer away. The experiment will send laser beams down 40-meter-long beam tubes (*2*) under vacuum. One set of beam tubes is being housed in a bunker formerly used for particle beams; the other juts out into the countryside, ending at a blue shed that houses a mirror and focusing optics (*3*). Precise optical equipment (*4*) is used to focus and align the beams.

Photographs by Matt Siber

© 2014 Scientific American



THE LONGEST VIEW

To learn how the universe evolved over time, a space telescope gazes back to the earliest galaxies ever observed

IN THIS HUBBLE SPACE TELESCOPE IMAGE, the galaxies of the giant cluster Abell 2744 appear strewn through space like jewels on black velvet. Their light began its journey to Earth some 3.5 billion years ago, when our biosphere was in its infancy. Despite appearances, the galaxies constitute less than 5 percent of the cluster's total mass, most of which resides in halos of invisible dark matter. There is, in fact, so much mass within this cluster that it behaves as a gravitational lens, warping space to magnify light from even deeper cosmic realms. The thousands of azure arcs and dots that surround the bright galaxies are actually distorted images of much fainter background galaxies, which were Hubble's true quarry in this image. We see them as they were more than 12 billion years ago, shortly after the big bang. They are some of the earliest galaxies ever glimpsed and offer a new window into galactic evolution, as well as a preview of what NASA's James Webb Space Telescope will routinely see after it launches later this decade. — *Lee Billings*

NASA, ESA AND J. LOTZ, M. MOUNTAIN, A. KOEKEMOER, AND THE HFF TEAM STScI

© 2014 Scientific American

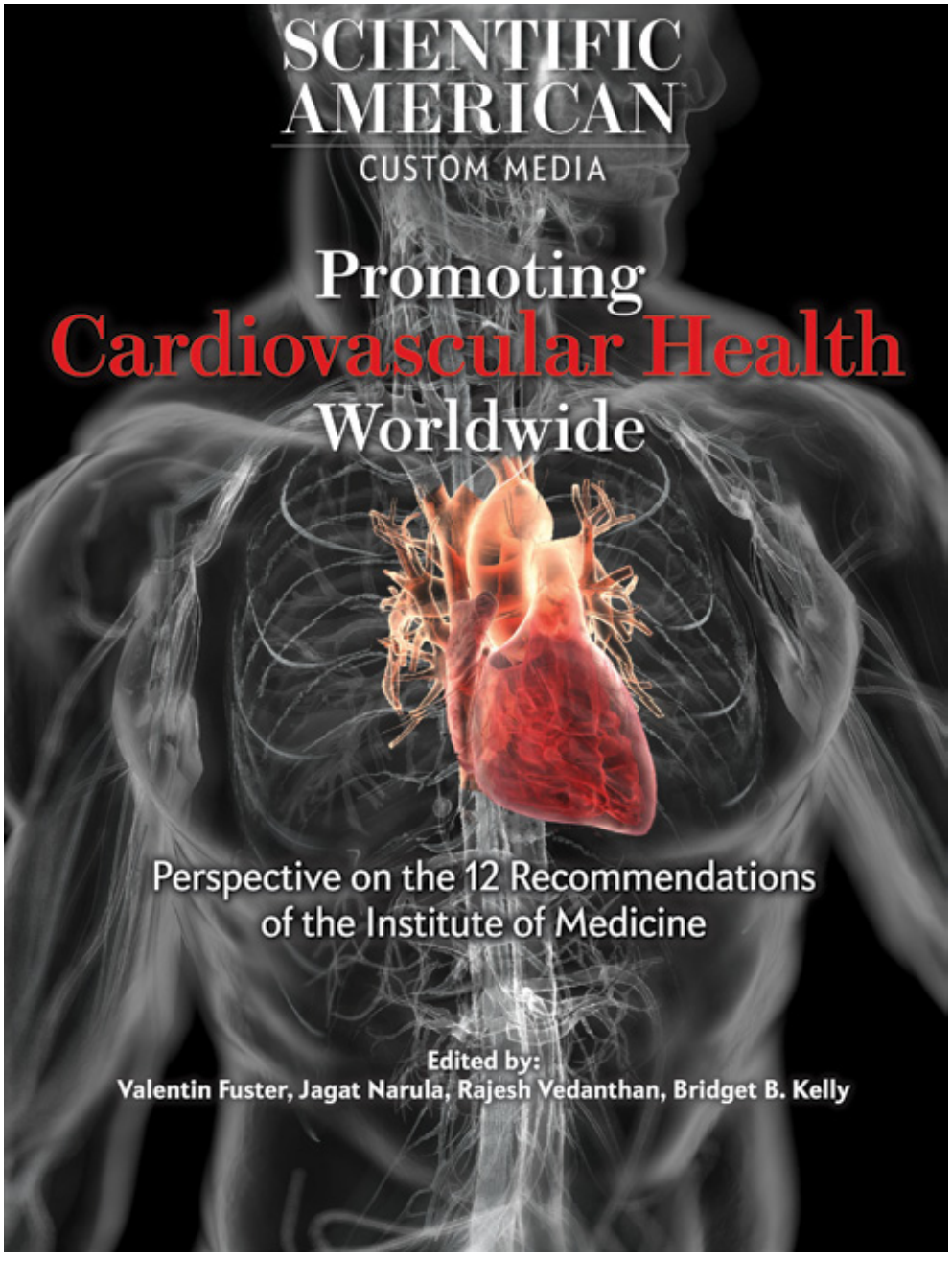
PROMOTING **CARDIOVASCULAR** HEALTH WORLDWIDE

A compelling look at one of the most pressing issues of our time...

Heart disease and related non-communicable diseases were once considered a problem that only wealthy, industrialized nations faced. Together, they now rank as the leading cause of death across the globe.

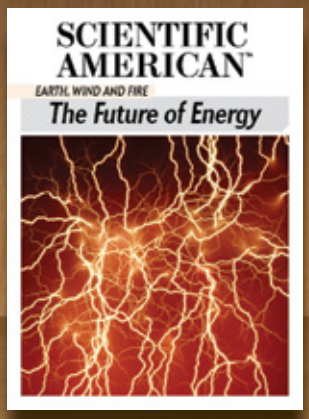
This special issue, ***Promoting Cardiovascular Health Worldwide***, explores the problem – and outlines the solutions.

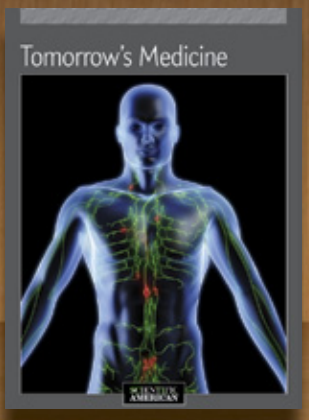
Brought to you by
SCIENTIFIC AMERICAN
CUSTOM MEDIA and
ICAHN SCHOOL
OF MEDICINE
AT MOUNT SINAI —
this special issue
is a follow up on the
12 Recommendations
from the Institute of
Medicine for promoting
cardiovascular health
in the developing world.

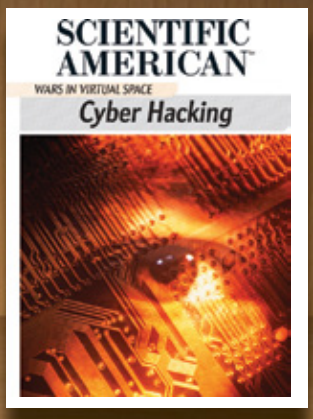


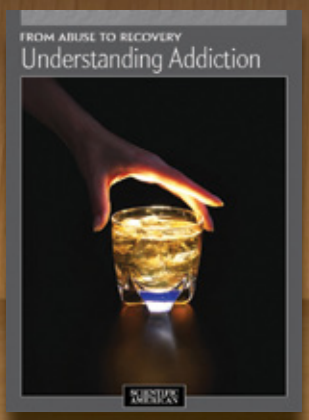
For a free digital issue visit:
ScientificAmerican.com/products/cardiovascular-health

eBooks From the Editors of SCIENTIFIC AMERICAN™









™ Scientific American, Inc.

Start your SCIENTIFIC AMERICAN eLibrary today:

books.scientificamerican.com/sa-ebooks